|  |
| --- |
| **FACULTAD DE INGENIERÍA**  **ESCUELA DE INGENIERÍA INFORMÁTICA** |
| **Asistente inteligente para la toma de decisiones durante el proceso de análisis de datos** |

**TRABAJO ESPECIAL DE GRADO**  
Presentado ante la   
**UNIVERSIDAD CATÓLICA ANDRÉS BELLO**Como parte de los requisitos para optar al título de   
**INGENIERO EN INFORMÁTICA**

Realizado por: Br. Miguel Obando

Profesor Guía: Ing. Livia Borjas

Fecha: Mayo, 2021

# Agradecimientos

Por qué me podría olvidar de muchos, porque nunca somos capaces de reconocer todo el bien que se nos ha hecho, por qué mis palabras a veces son torpes para expresar las cosas correctamente y por qué era lo que sentía que debía decir en este momento, dirijo mi agradecimiento: A todos.

**Tabla de contenido**

[Agradecimientos ii](#_Toc72324585)

[Índice de tablas vii](#_Toc72324586)

[Índice de figuras viii](#_Toc72324587)

[Capítulo I: 3](#_Toc72324588)

[El problema 3](#_Toc72324589)

[Planteamiento del problema 3](#_Toc72324590)

[Objetivo general 6](#_Toc72324591)

[Objetivo específicos 6](#_Toc72324592)

[Alcance 7](#_Toc72324593)

[Limitaciones 7](#_Toc72324594)

[Justificación 8](#_Toc72324595)

[Capitulo II 9](#_Toc72324596)

[Marco teórico 9](#_Toc72324597)

[Antecedentes de la Investigación 9](#_Toc72324598)

[DMDSS: Data Mining Based Decision Support System to Integrate Data Mining and Decision Support 9](#_Toc72324599)

[A Survey of Intelligent Assistants for Data Analysis 9](#_Toc72324600)

[Multi-agent based decision Support System using Data Mining and Case Based Reasoning 10](#_Toc72324601)

[Bases teóricas 11](#_Toc72324602)

[Dato 11](#_Toc72324603)

[Análisis de datos 12](#_Toc72324604)

[Análisis inteligente de datos 12](#_Toc72324605)

[Estadística 12](#_Toc72324606)

[Inteligencia Artificial 13](#_Toc72324607)

[Aprendizaje automático 13](#_Toc72324608)

[Knowledge Discovery o Descubrimiento de conocimiento 13](#_Toc72324609)

[Entender el dominio de aplicación y conocimiento relevante 14](#_Toc72324610)

[Recolección de datos 15](#_Toc72324611)

[Pre procesamiento de los datos 15](#_Toc72324612)

[Algoritmos de minería de datos y análisis de datos 25](#_Toc72324613)

[Visualización e interpretación de los resultados obtenidos 32](#_Toc72324614)

[Sistemas de soporte inteligentes 32](#_Toc72324615)

[Terminología básica 35](#_Toc72324616)

[Vista minable 35](#_Toc72324617)

[Experiencia de desarrollador 35](#_Toc72324618)

[Capitulo III 36](#_Toc72324619)

[Marco Metodológico 36](#_Toc72324620)

[Tipo de investigación 36](#_Toc72324621)

[Técnicas de recolección de datos 36](#_Toc72324622)

[Metodología de desarrollo 36](#_Toc72324623)

[Procedimiento metodológico 38](#_Toc72324624)

[Revisar el estado del arte de la construcción de asistentes inteligentes y de análisis de datos 38](#_Toc72324625)

[Definir un marco de trabajo para el análisis de datos basándose en el estado del arte previamente revisado 38](#_Toc72324626)

[Construir un asistente inteligente tomando como base la definición propuesta 41](#_Toc72324627)

[Validar el marco de trabajo implementado en el asistente inteligente mediante un caso de estudio 41](#_Toc72324628)

[Capitulo IV 43](#_Toc72324629)

[Desarrollo y resultados 43](#_Toc72324630)

[Revisión del estado del arte 43](#_Toc72324631)

[Definición del marco de trabajo 45](#_Toc72324632)

[Definición de criterios para los problemas de análisis de datos 47](#_Toc72324633)

[Construcción de asistente inteligente tomando como base la definición propuesta 53](#_Toc72324634)

[Criterios para selección de motor de reglas 53](#_Toc72324635)

[Drools 54](#_Toc72324636)

[Json-rules-engine 54](#_Toc72324637)

[Algoritmo de reglas basándose en json-rules-engine 54](#_Toc72324638)

[Interfaz 57](#_Toc72324639)

[Arquitectura del asistente 57](#_Toc72324640)

[Validación de marco de trabajo implementado 58](#_Toc72324641)

[Primer caso de estudio sin asistente 59](#_Toc72324642)

[Primer caso de estudio con asistente 60](#_Toc72324643)

[Segundo caso de estudio sin asistente 62](#_Toc72324644)

[Segundo caso de estudio con asistente 64](#_Toc72324645)

[Capítulo V 67](#_Toc72324646)

[Conclusiones y Recomendaciones 67](#_Toc72324647)

[Conclusiones 67](#_Toc72324648)

[Recomendaciones 67](#_Toc72324649)

[Bibliografía 69](#_Toc72324650)

[Anexos 75](#_Toc72324651)

# Índice de tablas

[Tabla 1 Ejemplo de datos 20](file:///C:\Users\ciber\Documents\TESIS\Trabajo%20de%20grado\Tesis%20en%20progreso%20Miguel%20Obando%20-%204-05.docx#_Toc71111449)

[Tabla 2 Ejemplo de Prueba-Resultado-Procedimiento 38](#_Toc71111450)

[Tabla 3 Tareas entregables y actividades de entendimiento del negocio 48](#_Toc71111451)

[Tabla 4 Reglas basándose en conocimiento adquirido de entendimiento del negocio 49](#_Toc71111452)

[Tabla 5 Tareas entregables y actividades de exploración de los datos 54](#_Toc71111453)

[Tabla 6 Reglas basándose en conocimiento adquirido de exploración de los datos 56](#_Toc71111454)

[Tabla 7 Tareas entregables y actividades de pre procesamiento por K medias 59](#_Toc71111455)

[Tabla 8 Reglas basándose en conocimiento adquirido para pre procesamiento de datos con algoritmo K medias 61](#_Toc71111456)

[Tabla 9 Tareas, entregables y actividades para pre procesamiento sobre algoritmo a priori 66](#_Toc71111457)

[Tabla 10 Reglas basándose en conocimiento adquirido de pre procesamiento para algoritmo a priori 67](#_Toc71111458)

[Tabla 11 Reglas de aplicación de algoritmo 69](#_Toc71111459)

[Tabla 12 Resultados de algoritmo de K medias sin pre procesamiento 75](#_Toc71111460)

[Tabla 13 Resultados de K medias siguiendo las recomendaciones del asistente 78](#_Toc71111461)

# Índice de figuras

[Figura 1 Ejemplo de integración: unificación de formatos 17](#_Toc72324655)

[Figura 2 Ejemplo de regla 40](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324656)

[Figura 3 Arquitectura de asistente inteligente. Adaptado de “Expert systems: Design and development” (1994) por Durkin J. (p. 170). 47](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324657)

[Figura 4 Esquema para creación de reglas 55](#_Toc72324658)

[Figura 5 Algoritmo de asistente para cumplimiento de reglas utilizando el esquema planteado 57](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324659)

[Figura 6 *Esquema de arquitectura de asistente inteligente* 58](#_Toc72324660)

[Figura 7 Muestra de datos académicos sobre los votos de los representantes del congreso de EE.UU para el año 1986 60](#_Toc72324661)

[Figura 8 Resultados de algoritmo a priori sin asistente 60](#_Toc72324662)

[*Figura 9* Atributos seleccionados por el clasificador 62](#_Toc72324663)

[Figura 10 Reglas seleccionadas por el algoritmo a priori con ayuda del asistente 62](#_Toc72324664)

[Figura 11 Muestra de datos de vidrios cargados al entorno WEKA 63](#_Toc72324665)

[Figura 12 Resultados obtenidos de aplicación de algoritmo de K medias sin ayuda del asistente 64](#_Toc72324666)

[Figura 13 Caso estudio con atributos reducidos 65](#_Toc72324667)

[Figura 14 Resultados de aplicación de algoritmo de k medias después de pre procesamiento con ayuda del asistente 66](#_Toc72324668)

# Índice de anexos

[Anexo 1 Tareas, entregables y actividades de entendimiento de negocio 75](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324717)

[Anexo 2 Reglas basadas conocimiento adquirido de entendimiento de negocio 77](#_Toc72324718)

[Anexo 3 Tareas, entregables y actividades de exploración de datos 80](#_Toc72324719)

[Anexo 4 Reglas basándose en conocimiento adquirido de exploración de los datos 82](#_Toc72324720)

[Anexo 5 Tareas entregables y actividades de pre procesamiento por K medias 84](#_Toc72324721)

[Anexo 6 Reglas basándose en conocimiento adquirido para pre procesamiento de datos con algoritmo K medias 86](#_Toc72324722)

[Anexo 7 Tareas, entregables y actividades para pre procesamiento sobre algoritmo a priori 88](#_Toc72324723)

[Anexo 8 Reglas basándose en conocimiento adquirido de pre procesamiento para algoritmo a priori 89](#_Toc72324724)

[Anexo 9 Reglas de aplicación de algoritmo a priori y k medias 91](#_Toc72324725)

[Anexo 10 Pregunta acerca de la disponibilidad de los datos realizada por el asistente durante etapa 1 del proceso de análisis de datos 92](#_Toc72324726)

[Anexo 11 Pregunta acerca de la definición de las fuentes de datos realizada por el asistente durante la etapa 1 del proceso de análisis de datos 93](#_Toc72324727)

[Anexo 12 Pregunta acerca de la organización realizada por el asistente durante la etapa 1 del proceso de análisis de datos 94](#_Toc72324728)

[Anexo 13 Recomendación del asistente acerca del tipo de técnica que debe ser utilizada en el proceso de análisis de datos 95](#_Toc72324729)

[Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos 96](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324730)

[Anexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos 96](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324731)

[Anexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori 96](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324732)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos 96](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324733)

[Anexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos 96](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324734)

[Anexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori 97](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324735)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori 97](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324736)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características 97](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324737)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori 97](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324738)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori 97](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324739)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características 98](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324740)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características 98](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324741)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características 98](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324742)

[Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características 98](file:///C:\Users\ciber\Desktop\respaldo%20tesis%202.docx#_Toc72324743)

[Anexo 17 Recomendación del asistente para utilizar técnicas de agrupamiento o clustering 99](#_Toc72324744)

[Anexo 18 Recomendación del asistente acerca de puntos a tener en cuenta para en análisis 100](#_Toc72324745)

[Anexo 19 Fragmento de recomendación del asistente acerca de cómo detectar valores anómalos y extremos 101](#_Toc72324746)

[Anexo 20 Fragmento de pantalla del asistente acerca de como eliminar valores anómalos 102](#_Toc72324747)

[Anexo 21 Fragmento de pantalla del asistente acerca de como aplicar el algoritmo de k medias 103](#_Toc72324748)

**Resumen**

En este trabajo de gradó se desarrolló un asistente inteligente para la toma de decisiones durante el proceso de análisis de datos basado en KDD. El estudio del arte acerca de la creación de los asistentes inteligentes así como el proceso de descubrimiento de conocimiento durante el desarrollo de este trabajo fue de vital importancia para posteriormente poder crear una base de conocimiento. Utilizando esta base y tomando como referencia los asistentes fundamentados en casos y reglas se desarrolló una estructura en el formato de notación de objetos de JavaScript que permite el encadenamiento de reglas hacia adelante, lo cual habilita que el asistente recabe información acerca del problema y brinde recomendaciones acerca de los métodos, filtros, herramientas e información que debe tener el o los responsables que llevan a cabo el proceso de análisis de datos. Mediante la validación del asistente con casos de estudio se pudo evidenciar la mejora de la precisión de los algoritmos que se tomaron en cuenta para la construcción de dicho asistente.

Palabras clave: Asistente inteligente, Análisis de datos, KDD

Debido a la relevancia que el proceso de análisis de datos ha tomado desde las últimas décadas en diversos ámbitos como lo son la medicina, las finanzas, y los negocios, muchas empresas han sumado a su plantilla personas encargadas de analizar los datos que sus sistemas recolectan para generar oportunidades de negocio u optimizar sus procesos.

La mayoría de los entornos para análisis de datos disponibles en el mercado están dirigidos a los expertos en análisis o requieren capacitación antes de ser útiles a los usuarios novatos. Sin embargo en la actualidad se requiere que los usuarios novatos sean capaces de llevar un proceso de análisis de datos para la organización con tan solo el conocimiento y experticia sobre el área donde se busca analizar dichos datos. En consecuencia la falta de experticia de los usuarios novatos les hace proclives a tomar dicciones no acertadas durante el proceso de análisis, producto de la abrumadora multitud de funciones y procedimientos disponibles, generando errores en los resultados, como por ejemplo, descubrir conocimiento poco relevante, no útil ni novedoso, interrupciones en el proceso debido a la incompatibilidad entre los datos y los procedimientos aplicados, entre otras inconsistencias.

Con el objetivo de crear asistencia para usuarios novatos durante los procesos de análisis de datos y así disminuir la cantidad de errores posibles, se desarrolla un asistente inteligente que permita guiar el usuario a través del proceso de análisis de datos bajo la metodología KDD apoyados en una herramienta de fácil acceso y extensiva documentación como lo es WEKA.

El asistente además, está desarrollado en el lenguaje de programación JavaScript y haciendo uso de archivos de tipo notación de objeto de este mismo lenguaje dentro de su base de conocimiento para que este pueda ser fácilmente extendido y pueda abarcar métodos y técnicas que no se contemplaron en la realización de este trabajo de grado.

El trabajo se organiza en cinco capítulos en los que se desarrolla lo siguiente: el capítulo I define el planteamiento del problema, objetivos, la solución propuesta y su justificación; el capítulo II explica toda la teoría necesaria para comprender los conceptos empleados en la solución así como conceptos clave conceptos clave para entender el proceso de análisis de datos; el capítulo III presenta y justifica las metodologías de investigación y desarrollo usadas como guía para lograr paso a paso los objetivos; el capítulo IV describe la implementación de la solución y los resultados que han sido obtenidos mediante el uso de un caso de estudio; el capítulo V menciona las conclusiones obtenidas en base a los objetivos planteados y las recomendaciones para futuros trabajos relacionados con el desarrollo de los asistentes inteligentes para los procesos de análisis de datos

# Capítulo I:

# El problema

## Planteamiento del problema

Debido a la relevancia que el proceso de análisis de datos ha tomado desde las últimas décadas en diversos ámbitos como lo son la medicina, las finanzas, y los negocios, muchas empresas han sumado a su plantilla personas encargadas de analizar los datos que sus sistemas recolectan para generar oportunidades de negocio u optimizar sus procesos (Resagratia, 2020). Análisis de datos se define como la descomposición, observación y estudio de un conjunto de datos con el objetivo de responder preguntas acerca de a estos (Berthold, 2007, págs. 2-3). Su importancia recae en que las organizaciones, mediante este proceso, pueden descubrir modelos entre los datos, inadvertidos y no implícitos, de mucho interés que redundan en diversidad de beneficios, tales como: conocer a sus clientes, conocer las futuras tendencias de consumo, disminuir costos de operación, entre otros beneficios (Hinshelwood, 2008). Un ejemplo poco conocido es el club de futbol inglés Liverpool, que logró ganar la liga de campeones del año 2020 con la ayuda de un equipo de analistas de datos, lo que les permitió saber cuál era el mejor estilo de juego que se adaptaba a ellos y bajo qué condiciones, así como la probabilidad de anotar goles en ciertas posiciones de sus jugadores basándose en los datos de sus partidos anteriores (Puiu, 2020).

El proceso de analizar los datos para identificar patrones no triviales, novedosos y potencialmente útiles es llamado descubrimiento de conocimiento (Fayyad , Smyth, & Piatetsky-Shapiro, 1996, págs. 82-88). Este análisis de datos puede ser de dos tipos, descriptivo o predictivo. El análisis descriptivo tiene por objetivo describir un conjunto de datos mediante la agrupamiento de los mismos en grupos llamados predictores categóricos, mientras que el predictivo busca inferir valores que no se conocen tomando en cuenta el conjunto dado (Berthold, 2007, pág. 4).

El descubrimiento de conocimiento, conocido como KDD (Knowledge Discovery in Databases en inglés) es un proceso de cinco fases: 1) entender el dominio de aplicación y conocimiento relevante del problema, 2) seleccionar los datos relevantes para el problema a resolver 3) pre procesar los datos seleccionados 4) Pasar los datos través de algoritmos de minería de datos y análisis de datos, 5) Mostrar e interpretar los resultados obtenidos (Fayyad , Smyth, & Piatetsky-Shapiro, 1996, pág. 82). Este proceso es un proceso interactivo e iterativo que involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016, págs. 63-86).

La mayoría de los entornos para análisis de datos disponibles en el mercado están dirigidos a los expertos en análisis o requieren capacitación antes de ser útiles a los usuarios novatos (Goebel & Gruenwald, 1999, pág. 31). Lo usuarios novatos (personas ajenas al área de análisis de datos) generalmente están menos capacitadas en el análisis de datos complejos pero suelen ser personas con mucha experticia y conocimiento valioso sobre el problema o pregunta que se busca responder con el análisis de datos (Goebel & Gruenwald, 1999, pág. 31). Además, normalmente no están interesados en usar tecnología avanzada y poderosa, solo en obtener respuestas claras y rápidas a sus preguntas cotidianas de negocios. (Goebel & Gruenwald, 1999, pág. 31).

Sin embargo en la actualidad se requiere que los usuarios novatos sean capaces de llevar un proceso de análisis de datos que genere conocimiento valioso para la organización con tan solo el conocimiento y experticia sobre el área donde se busca analizar los datos; sin embargo, muy pocas herramientas del mercado poseen asistencia guiada para llevar a cabo dicho proceso (Goebel & Gruenwald, 1999, pág. 31). Esto hace que los usuarios novatos tengan errores en los resultados, interrupciones en sus procesos de análisis debido a la incompatibilidad entre los datos y los procedimientos aplicados, o información poco relevante debido a la abrumadora multitud de funciones y procedimientos disponibles (Serban, Vanschoren, Kietz, & Bernstein, 2013, pág. 2).

Una de las consecuencias de la falta de asistencia guiada a los usuarios novatos es la mala preparación de los datos, lo que conlleva a un rendimiento menos eficiente de los modelos utilizados en comparación con aquellos modelos con la preparación adecuada de sus datos (Wielenga, 2007, pág. 1).

Otras de las consecuencias de la falta de asistencia para los usuarios novatos en un proceso de análisis de datos es que a menudo este tipo de usuario busca los datos o variables mínimas requeridas para que un modelo arroje resultados (Wielenga, 2007, pág. 1). Sin embargo, las organizaciones a pesar de tener datos similares, frecuentemente tienen un número razonablemente grande de detalles que recogen y que sus competidores no. Adicional a esto la frecuencia con que estos se almacenan puede diferir. Si todas las organizaciones utilizan el mismo subconjunto de datos se pierden oportunidades de identificar patrones únicos y deja que los competidores que utilizan toda la riqueza de sus datos incrementen las posibilidades de ganar cuota de mercado con el paso del tiempo (Wielenga, 2007, pág. 1). Este problema mediante asistencia guiada podría adaptar el modelo a utilizar (de acuerdo a los datos cargados) para obtener un mejor rendimiento del mismo.

Otro de los problemas derivado de la falta de asistencia guiada en el proceso de análisis de datos que ocurre, es la preparación incorrecta de los predictores categóricos (o grupos de categoría) cuando se realiza un modelado de categorización (Wielenga, 2007, pág. 2). En este sentido suelen presentarse las siguientes situaciones: 1) “tener demasiados niveles generales o categorías”, lo cual genera problemas de rendimiento en los algoritmos, 2) “tener niveles o categorías que ocurren raramente”, lo que causa poco impacto en el modelo y 3) “tener niveles o categorías con características que ocurren frecuentemente”, lo cual hace que un nivel explique todo el conjunto de datos analizado, generando en consecuencia información no relevante (Wielenga, 2007, pág. 2). Este problema mediante asistencia guiada en el criterio de la selección de predictores podría optimizar los creados por el usuario.

Con base a lo expuesto anteriormente se propone el desarrollo de un asistente inteligente (un sistema que tomando en cuenta los datos y el problema a resolver recomiende la decisión más óptima) que guíe al usuario novato en la toma de decisiones durante un proceso de análisis de datos. Este asistente ayudará a reducir la cantidad de errores más comunes que comenten los usuarios novatos al momento de utilizar un entorno de análisis de datos.

Un asistente inteligente (también conocido como sistema de soporte de decisiones aportará al problema de asistir de manera inteligente al proceso de análisis de datos las siguientes prestaciones: reducirá la incertidumbre en la toma de decisiones en los proceso de análisis de datos (Chen, 2004, pág. 104), incrementará el rendimiento de las herramientas de toma de decisiones (Rupnik, Kukar, Bajec, & Krisper, 2006, págs. 225-230), mejorará y ayudará a crear distintos enfoques que se tienen de los problemas actuales que buscan el análisis de datos como solución y permitirá fusionar el conocimiento de los expertos en análisis de datos con el conocimiento que tienen los usuarios novatos en su área.

## Objetivo general

Desarrollar un asistente inteligente para la toma de decisiones en el proceso deanálisis de datos.

## Objetivo específicos

1. Revisar el estado del arte de la construcción de asistentes inteligentes y de análisis de datos.
2. Definir un marco de trabajo para el análisis de datos basándose en el estado del arte previamente revisado.
3. Construir un asistente inteligente tomando como base la definición propuesta.
4. Validar el marco de trabajo implementado en el asistente inteligente mediante un caso de estudio.

## Alcance

En un proceso de KDD la fase de pre procesamiento de datos es la más compleja y tiene variantes importantes a considerar (Knight, 2015), de tal manera que su ejecución generalmente requiere la mayor inversión de recursos y así como el 50-80% del tiempo de todo el proceso de análisis.

Tomando en consideración la importancia de la fase de pre procesamiento de datos y su complejidad, así como gestionando el factor tiempo en la planificación del presente trabajo de grado, se define como alcance desarrollar un asistente inteligente para soportar la toma de decisiones durante las cuatro (4) primeras fases del proceso de KDD, de tal manera que el asistente brindará soporte desde el pre procesamiento de la data hasta la fase de aplicación de la minería de datos.

## Limitaciones

El asistente inteligente estará configurado sobre un ambiente de trabajo de código abierto y licencia gratuita. Adicionalmente, aunque el asistente pasará por las cuatro primeras fases de un proceso de descubrimiento, brindará recomendaciones en la toma de decisiones del usuario en el proceso de análisis de datos hasta generar una vista minable y la selección y aplicación de los algoritmos de minería de datos, ayudando a remover el ruido de los datos, los valores atípicos si corresponde, recolectar la información necesaria para modelar o llevar registro del ruido, apoyar y sugerir estrategias para manejar los valores faltantes en los registros, y sugerir posibles soluciones a problemas relacionados a los repositorios de datos donde se extrae la información. Todo lo anterior tomando en cuenta la retroalimentación que el usuario le provea.

Finalmente, el asistente se limitará a dar recomendaciones en procesos de descubrimiento de conocimiento sobre repositorios de datos y, de acuerdo al entorno seleccionado, el asistente no utilizará todas las posibles herramientas y funcionalidades disponibles en el mercado, más allá de las que el entorno ofrezca por defecto además de no hacer uso de herramientas para el análisis de datos sobre imágenes.

## Justificación

El presente proyecto redundará en diversos beneficios, entre los que se destacan: facilitar al usuario el uso de las herramientas disponibles en un entorno especializado en descubrimiento de conocimiento, así como el contexto en que estas pueden o deben ser utilizadas.

El asistente inteligente guiará al usuario novato a través del proceso de análisis de datos, mostrando las diversas herramientas disponibles y los posibles procedimientos que pueden seguir sus datos de acuerdo a la retroalimentación que el usuario le suministre, permitiendo mejorar el criterio de las decisiones a tomar en un proceso de descubrimiento.

# Capitulo II

# Marco teórico

## Antecedentes de la Investigación

Para la realización de este trabajo de grado se tomará en cuenta los siguientes trabajos, con el uso de asistentes inteligentes y afines:

### DMDSS: Data Mining Based Decision Support System to Integrate Data Mining and Decision Support

En este trabajo sus autores explican el proceso de la implementación de un sistema de soporte basado en decisiones como herramienta para permitir a los usuarios novatos utilizar técnicas de minería de datos (una parte vital de proceso de descubrimiento de conocimiento) a usuarios con poca experiencia en el área de descubrimiento de conocimiento (Rupnik, Kukar, Bajec, & Krisper, 2006).

Los autores destacan que solo con conocimiento básico acerca de minería de datos y con la ayuda del sistema de soporte los usuarios pudieron descubrir las ventajas del uso continuo de analizar sus datos. Los usuarios aplicaron mejoras al área de mercadeo basándose en la información obtenida de los modelos que prepararon con ayuda del sistema. Adicionalmente, meses después los usuarios entendieron el verdadero potencial de esta herramienta y entendieron de forma profunda cómo funcionan los modelos que el sistema recomendó, lo que permitió obtener nuevas ideas y enfoques en su análisis de datos para obtener resultados significativos

### A Survey of Intelligent Assistants for Data Analysis

En este trabajo sus autores analizan el estado de los asistentes inteligentes para la toma de decisiones en un proceso de descubrimiento de conocimiento hasta la fecha, destacando algunos entornos como lo son IBM SPSS Modeler, RapidMiner, SAS Enterprise Miner y WEKA (Serban, Vanschoren, Kietz, & Bernstein, 2013).Adicionalmente señalan que estos entornos a pesar de ser muy flexibles y poseer una gran cantidad de operaciones sobre los dato, no son amigables para el usuario novato (Serban, Vanschoren, Kietz, & Bernstein, 2013).

Adicionalmente sus autores varían analizan que tipo de conocimiento el usuario debe contar para que el asistente pueda ser de utilidad, los asistentes disponibles en el mercado actualmente, limitaciones de cada uno de ellos y por ultimo recomendaciones o puntos de mejora a nivel general para este tipo de sistemas.

Los autores concluyen y resaltan que a medida que la exploración de datos se va haciendo importante en las configuraciones industriales y científicas, la necesidad de soporte automatizado para análisis de datos probablemente incrementará (Serban, Vanschoren, Kietz, & Bernstein, 2013).

### Multi-agent based decision Support System using Data Mining and Case Based Reasoning

En este trabajo, sus autores presentan las propuestas de crear un modelo integrando sistema de soporte de decisiones, análisis de datos, sistemas de razonamiento basado en casos y sistemas multi-agentes, lo cual permitiría desarrollar un sistema de apoyo de decisiones bajo situaciones complejas, que de esta manera otra de las formas en que el proceso de descubrimiento de conocimiento puede ser apoyado (Srinivasan, Singh, & Kumar, 2011) .

Entre las ventajas de utilizar agentes inteligentes en el campo de análisis de datos y los sistemas de apoyo de decisiones destacan la agilización de los procesos para llegar a los objetivos por los cuales son implementados en las organizaciones, así como buscar y filtrar entre grandes cantidades de información. La combinación de distintos agentes permitirían volverlos adaptable a cualquier situación, así como aumentar la robustez de las decisiones sugeridas por los mismos debido a la colaboración y especialización de cada uno de los mismos.

Adicional a esto, los autores destacan que los sistemas de toma de decisiones deben tener la habilidad de aprender y adaptarse a las condiciones de los ambientes en los que se desarrollan debido a la volatilidad de los mercados.

Estos trabajos fueron escogidos como antecedentes de la investigación por relevancia acerca de cómo los asistentes inteligentes han sido implementados así como las tecnologías utilizadas. Adicionalmente muestran de manera concisa como debe ser la estructura base de la toma de decisiones de cualquier asistente inteligente así como la implementación sobre un proceso de descubrimiento de conocimiento y los factores que deben ser tomados en cuenta.

## Bases teóricas

Para el correcto entendimiento de la presente investigación es importante tener claridad sobre ciertas definiciones básicas sobre las cuales será desarrollado. En esta sección se presentan distintos conceptos con el detalle necesario para el desarrollo del trabajo, partiendo con la definición de dato, el análisis de los mismos, los métodos utilizados para su análisis y posteriormente el proceso de descubrimiento de conocimiento y cómo los sistemas de soporte inteligentes colaboran con dicho proceso.

### Dato

Se entiende como dato como “información sobre algo concreto que permita su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho” (Real Academia Española, 2020) y adicionalmente la RAE en el ámbito informático lo define como “información dispuesta de manera adecuada para su tratamiento por una computadora”. En la actualidad la recolección y análisis de los datos es un tema de mucho interés muy relacionado con la evolución de la tecnología que los captura y gestiona.

### Análisis de datos

El análisis de datos se define como la descomposición, observación y estudio de un conjunto de datos con el objetivo de responder preguntas acerca de los mismos (Berthold, 2007, págs. 2-3).

El análisis de datos se divide en dos tipos, el descriptivo y el predictivo. El análisis descriptivo tiene por objeto realizar afirmaciones sobre un conjunto de datos, por ejemplo ¿Cuántas mujeres existen en esta población? ¿Cuál es la característica más común de estos productos? entre otros (Berthold, 2007, pág. 4). Por otro lado el análisis predictivo busca generar conclusiones con una validez más general, en donde todo el conjunto de datos no ha sido observado, por ejemplo ¿el número de autos aumentará para el próximo año? ¿El número de mujeres se mantendrá en determinada localidad durante estos próximos 3 meses? (Berthold, 2007, pág. 4).

### Análisis inteligente de datos

Por otro lado se le llama análisis inteligente de datos cuando se aplican herramientas estadísticas y de aprendizaje automático para ayudar a responder las preguntas sobre un conjunto de datos dado en un proceso de análisis de los mismos (Berthold, 2007, pág. 4).

### Estadística

La estadística es “una ciencia formal que estudia la recolección, análisis e interpretación de datos de una muestra representativa, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado” (Depool & Monasterio, 2013).

### Inteligencia Artificial

Inteligencia artificial es “la teoría y el desarrollo de sistemas computacionales capaces de desempeñar tareas que normalmente requieren inteligencia humana, como percepción visual, reconocimiento de voz, toma de decisiones y traducción entre lenguajes” (Wright & Daintith, 2008)

### Aprendizaje automático

El aprendizaje automático por su parte es la rama de la inteligencia artificial que se encarga de dar la capacidad a las programas de computadoras de aprender (o mejorar) tareas basado en los datos recolectados (Witten, Frank, & Hall, Data mining: Practical machine learning tools and techniques, 2011, pág. 8). De acuerdo a (Han, Kamber, & Pei , 2012, pág. 24). “un ejemplo típico de aprendizaje automático es programar una computadora para que ésta pueda reconocer códigos postales escritos a mano en un correo después de haber aprendido de un conjunto de ejemplos”.

Existe una continuidad entre aprendizaje automático y estadística en donde la estadística ha ayudado a desarrollar la rama del aprendizaje automático y viceversa. La principal diferencia entre ambas ramas es que la estadística se interesa más en probar hipótesis mientras que el aprendizaje automático se centra más en generar procesos de generalización como un paso en la búsqueda de una posible hipótesis (Witten, Hall, Frank, & Pal, 2017, págs. 28-29). A pesar de lo anteriormente dicho, muchísimas técnicas estándar de estadística han sido adaptadas por investigadores de aprendizaje automático (Witten, Hall, Frank, & Pal, 2017, pág. 29).

### Knowledge Discovery o Descubrimiento de conocimiento

Tanto las técnicas de aprendizaje automático como las técnicas estadísticas son ampliamente utilizadas para el análisis de datos, siendo un amplio espectro de técnicas a considerar. Sin embargo el presente trabajo se enfocará en el proceso iterativo de descubrimiento de conocimiento, que depende de la toma de decisiones de forma dinámica y se adecua a múltiples contextos, lo que permite la intervención de los asistentes inteligentes como guías durante el proceso

El descubrimiento de conocimiento (también llamado KDD Knowledge Discovery in Databases, por sus siglas en inglés) es un proceso que tiene las siguientes etapas: 1) entender el dominio de aplicación y conocimiento relevante, 2) seleccionar los datos relevantes para el problema a resolver, 3) pre procesar (limpiar y transformar) los datos seleccionados e incorporar el conocimiento previo que se tiene del problema, 4) pasar los datos previamente transformados a través de algoritmos de minería de datos y análisis de datos, 5) mostrar e interpretar los resultados obtenidos (Fayyad , Smyth, & Piatetsky-Shapiro, 1996, págs. 82-88). Esto con el fin de convertir los datos en conocimiento útil para la organización que lleva a cabo dicho procedimiento.

El asistente inteligente desarrollado en el presente trabajo servirá como apoyo desde la etapa de pre procesamiento de datos hasta la etapa de minería de datos, etapas 1 hasta 4 previamente citadas. A continuación se definen cada uno de los pasos (y métodos relacionados a estos) del KDD por los cuales el asistente brindará soporte, destacando primero en que consiste la etapa para que luego (una vez el lector tenga contexto al respecto) definir en que apoyará el asistente:

### Entender el dominio de aplicación y conocimiento relevante

La mayoría de los estudios de modelado basados en datos son desarrollados en un dominio de aplicación particular (Kantardzic, 2020, pág. 7). Por lo tanto, la experiencia y el conocimiento sobre el campo al donde se desea hacer el estudio generalmente es requerido para poder hacer una declaración del problema detallada (Kantardzic, 2020, pág. 7). En este paso del KDD es necesario que el experto en el modelado y la persona que posee el conocimiento sobre el campo que se desee estudiar estén en constante para que las variables del problema cuya dependencia se desconoce pueda ser establecida y de ser posible establecer una hipótesis inicial respecto a dichas variables.

En esta etapa del proceso el asistente inteligente brindará ayuda mediante una cantidad de preguntas que guiaran al usuario a definir el objetivo de su análisis.

### Recolección de datos

Dentro de la recolección de datos de un proceso de KDD existen 2 enfoques, el primero es llamado “experimento diseñado” en el cual el proceso de generación de datos está controlado por el encargado del proceso de KDD y el segundo enfoque es cuando el encargado del proceso de KDD no influye de ninguna manera en la recolección de los mismos, este es llamado “enfoque observacional” (Kantardzic, 2020, págs. 7-8).

En esta etapa, el asistente inteligente, tomando en cuenta las respuestas entregadas por el usuario, brindará soporte al usuario acerca de la cantidad y el tipo de datos que debe tomar.

### Pre procesamiento de los datos

La limpieza de datos o pre procesamiento de datos se define como el proceso de eliminar errores e inconsistencias de los datos, así como resolver el problema de identidad del objeto (Galhardas, Florescu, Sasha, Simon, & Saita, 2001), ya que generalmente alrededor de un 5% de los datos o más tienden a tener errores (Orr, 1998, págs. 66-71), este rango de error puede llegar a un 40%. (Fayyad, Piatetsky-Shapiro, & Uthurasamy, Data Mining: The Next 10 Years, 2013)

El pre procesamiento de datos, puede derivar en las siguientes fases que no necesariamente se deben realizar en el orden mostrado:

* Integración de datos
* Reconocimiento
* Valores faltantes
* Detección de valores anómalos
* Transformación de atributos
* Reducción de dimensionalidad por transformación
* Aumento de dimensionalidad por transformación o construcción
* Discretización
* Numeración
* Normalización de rango

Para cada una de estas se utilizan técnicas estadísticas de visualización y de consulta o incluso técnicas más artesanales y variadas como lo son redefinición de atributos, discretización y numeración.

#### Integración de datos

A menudo durante la recolección de datos se extraen datos de distintas fuentes y pasan a unirse, por lo que de acuerdo a Orallo J., Ramírez Quintana M. J & Ferri Ramírez C. (2004 pág. 68*)* ocurren los siguientes problemas

* Dos o más fuentes de objetos diferentes se unifican: por lo que los datos resultandos mezclan características de distintos objetos y pasan a generar resultados erróneos al momento de extraer conocimiento, este problema se agudiza cuando la diferencia entre los objetos es más grande
* Dos o más fuentes de objetos iguales se dejan separadas: las características de un mismo objeto aparecen repartidas entre varios objetos parciales, esto es especialmente problemático cuando se usan valores agregados, por ejemplo, el total de compras será mayor si consideramos 1 objeto como 2 separados dentro de una base de datos

*“*En general primer problema es menos frecuente que el segundo ya que la unificación generalmente por identificadores externos a la base de datos” (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 68) sin embargo, esta tarea puede ser más compleja si se utilizan claves internas para identificar objetos (por ejemplo, auto enumeradas), debido a que generalmente varían de formato (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 68).

Otro caso muy frecuente es “la integración de formatos diferentes, que se produce si tenemos codificaciones diferentes, idiomas diferentes, medidas diferentes, entre otros” (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 68). Un ejemplo de esto es cuando se extraen datos de países que utilizan sistemas de medidas distintos, como por ejemplo, el kilogramo y la libra. En la figura 1 se puede observar un ejemplo de integración

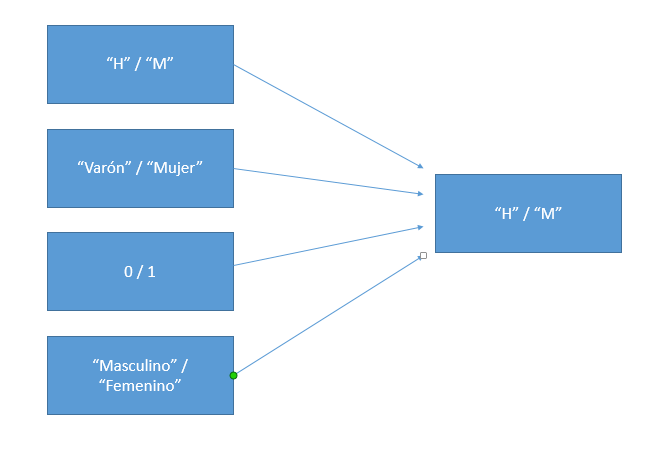


Figura 1 Ejemplo de integración: unificación de formatos

#### Reconocimiento

Una vez se tienen todos los datos integrados, se puede realizar una identificación mediante un resumen de características de los atributos, ya sea de manera individual para cada una de las tablas generadas o para toda la base de datos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 69).

De esta manera por cada uno de los atributos, obteniendo la cantidad total de datos, la media, la desviación estándar, la moda, y los valores máximos y mínimos podemos validar y descubrir datos inconsistentes o ambiguos, ejemplo, cuando en un atributo sexo se tienen más de 2 valores como lo son “H”, “V” y “M”, se debe recurrir a las fuentes de donde se obtuvieron los datos y al dominio del problema para eliminar ambigüedades, ya que, “H” puede significar “Hombre” o “Hembra” y “M” puede significar “Mujer” o “Masculino”.

#### Valores faltantes

Los valores faltantes de los atributos de los elementos pueden ser remplazados por varias razones, el método de minería de datos con que se busca extraer conocimiento puede verse afectado de manera negativa por la falta la omisión de estos, en segundo lugar, podemos querer agregar los datos (especialmente los numéricos) para realizar otro tipo de operaciones con los mismos y que estos valores sean necesarios y en tercer lugar es posible que de minería de datos ignore todo el individuo debido a la falta de este atributo (produciendo así un sesgo) o que el método de sustitución de los valores faltantes no sea el adecuado debido a que no conoce el contexto del mismo (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 74).

La detección de estos valores no siempre puede ser sencilla a simple vista debido a que aunque muchas veces se dejan como vacío en ocasiones los sistemas conectados al almacén de datos tienden a hacer obligatorios algunos campos, por lo que el usuario coloca valores no vacíos pero no validos como lo son “-” en una dirección, o “0000-0000-0000-0000-000” para una tarjeta de crédito por ejemplo (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 75).

Orallo J., Ramírez Quintana M. J y Ferri Ramírez C. (2008 pg. 75) resalta que es importante que se tome en cuenta el contexto tanto para la detección como su tratamiento posterior debido a que:

* Algunos valores faltantes explican características relevantes: tener números de teléfonos inválidos puede representar el deseo de los clientes de no ser molestados.
* Valores no existentes: muchos valores faltantes existen en la realidad, pero otros no, un ejemplo de esto es que un cliente que se acaba de registrar en el sistema no presente un historial crediticio.
* Datos incompletos: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer una unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una/s fuente/s diferente/s al resto.

Una vez identificados los datos faltantes y sus causas se pueden tomar las siguientes estrategias para lidiar con los datos faltantes:

* Borrar registros con datos faltantes: Este método consiste en ignorar todos los registros con casos incompletos, esto en estadística es llamado caso de análisis completo (Grzymala-Busse & Grzymala-Busse, 2005, pág. 35).
* Valor más común de un atributo: Esta estrategia consiste en remplazar los valores faltantes con el valor más común del atributo, obteniendo ese valor de la frecuencia relativa de los valores existentes de cada atributo (Grzymala-Busse & Grzymala-Busse, 2005, pág. 35) .
* Valor más común restringido a un sub conjunto de datos: Esta estrategia es similar a la mencionada anteriormente pero se limita la frecuencia de acuerdo a unos registros específicos, supongamos que tenemos los siguientes datos mostrados en la tabla 1

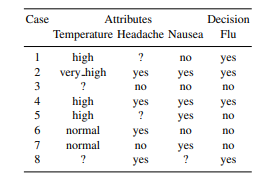


Tabla 1

Ejemplo de datos

Tomado de *Data mining and knowledge discovery handbook (pg. 34)* por *Grzymala*-Busse J.W., Grzymala-Busse W.J. 2005.

En este caso, si tomamos como concepto los atributos 1, 6 y 7, tendríamos entonces que el atributo “temperature” tiene como valor más frecuente “normal”, por lo que pasaríamos a remplazar los valores faltantes de ese atributo como “normal”.

* Asignar todos los posibles valores a los atributos con valores faltantes: Para esta estrategia, como su nombre lo indica, asigna todos los posibles valores
* Remplazar valores de atributos faltantes de acuerdo a la media de los valores del atributo: Esta estrategia está restringida a valores numéricos y consiste en obtener la media de los valores presentes para asignarle a los valores faltantes dicho resultado (Grzymala-Busse & Grzymala-Busse, 2005, pág. 39).
* Remplazar valores de atributos faltantes por la media de los valores de un atributo restringido a un subconjunto de datos: Esta estrategia está restringida al igual que el anterior está restringido a valores numéricos y consiste en obtener la media de acuerdo a un subconjunto de registros previamente definido (Grzymala-Busse & Grzymala-Busse, 2005, pág. 40).
* Ajuste global más cercano: El ajuste global más cercano consiste en remplazar el valor faltante con otro registro que sea lo más parecido a este, para este método se comparan todos los registros de forma vectorial tomando como pivote todos los demás datos que el registro con el atributo faltantes contiene, tomando así el atributo del registro que resulte con la distancia (Grzymala-Busse & Grzymala-Busse, 2005, págs. 40-41):
* Ajuste global restringido a un subconjunto de datos: Esta estrategia es similar al método anterior pero restringido a un subconjunto de datos previamente seleccionado (Grzymala-Busse & Grzymala-Busse, 2005, pág. 41).
* Otros métodos: Adicionalmente existen otra serie de métodos estadísticos para manejar los valores faltantes conocidos como imputación, así como métodos de aprendizaje automático basados en una serie de restricciones y árboles de decisiones (Grzymala-Busse & Grzymala-Busse, 2005, pág. 43).

#### Detección de anomalías

Marcus A. y Maleti Jonathan I. (2010 pág. 23) definen el proceso de limpieza de datos en 3 fases:

* Definir y determinar los tipos de errores.
* Buscar y determinar las instancias de los errores previamente definidos.
* Corregir errores no cubiertos.

Una vez definidos y determinados los errores se buscan determinar los datos anómalos, para esto existen 4 métodos que comúnmente son utilizados:

##### Método estadístico:

El método estadístico busca de acuerdo a Barnett y Lewis (1994) “identificar campos anómalos y registros utilizando valores como la media, la moda y la desviación estándar basados en el teorema de Chebyshev” considerando los intervalos de confianza de cada uno de los campos. Este método tiene la contra de poder generar muchos falsos positivos sin embargo, es rápido y simple, por lo que generalmente tiende a ser usado en combinación con otros métodos (Maletic & Marcus, 2010, pág. 44) .

Para la aplicación de este método se utiliza la media y la desviación estándar, para cada valor se toman los intervalos de confianza en donde para cada valor se utilizan intervalores de confianza definidos por:

(1)

En donde es el valor del registro, i es el índice del registro, es la media, es la desviación estándar y un valor definido por el usuario, basado en la experiencia del campo en el que se está haciendo el descubrimiento de conocimiento (Maletic & Marcus, 2010, pág. 24).

##### Agrupamiento

El método de agrupamiento consiste en utilizar algoritmos de distancia euclidiana sobre los datos, la mayor desventaja de este método es la alta complejidad computacional que requiere (Maletic & Marcus, 2010, pág. 25) .

##### Métodos basados en patrones

Un patrón se define como una colección de registros que tiene características o comportamientos similares basados en un porcentaje definido por el usuario (generalmente sobre el 90%), la identificación de los patrones pasa por la combinación de una serie de técnicas como lo son particionamiento, clasificación y agrupamiento (Maletic & Marcus, 2010, pág. 25).

##### Reglas de asociación

Las reglas de asociación definen distintos tipos de patrones, los registros que no siguen estos patrones son considerados anómalos, este proceso es similar a la detección de patrones, pero utilizando reglas ordinales, la ventaja de este método es que puede tratar con distintos tipos de datos (Maletic & Marcus, 2010, pág. 26)

#### Discretización y numeración

Uno de los aspectos más importantes dentro de un proceso de KDD es el tipo de los atributos de los datos, esto determina como va a ser tratado por el algoritmo que sea aplicado dentro de la fase posterior de minería de datos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 89).

##### Discretización

La discretización (también llamado en inglés como *binning*) consiste en convertir atributos numéricos a nominales ordenados, la discretización se utiliza en los casos en donde el error de medida puede ser grande o existen umbrales significativos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, págs. 89-90). Un ejemplo de esto es que se quisiera hacer un estudio sobre las notas de distintas universidades sobre cierta asignatura se deben llevar sus notas a una escala común, debido a que la escala entre universidades varía, en España por la nota de aprobación mínima es 5 y en El salvador es 6, por lo que es pertinente transformar todas estas notas a una escala común para el algoritmo a utilizar (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 90)

##### Numerización

La numerización es el proceso en el cual se le asignan atributos indicadores a los datos de acuerdo a una serie de condiciones basándose en una escala numérica, si quisiéramos por ejemplo separar una población entre niño, adolescente, adulto, anciano, podríamos hacerlo mediante una numerización tomando en cuenta el rango de edades.

##### Normalización de rango

Existen algunos casos especiales en donde es necesario normalizar rangos al mismo atributo, un ejemplo de esto es cuando el mismo atributo viene de distintas fuentes (euros, pesetas) o cuando se tiene planteado trabajar con algoritmos que trabajan con distancias ya que por ejemplo, los atributos que van de 0 a 100 serán mucho mayores que los que van de 0 a 10 (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 93). La normalización más común para estos casos es llamada normalización normal uniforme y normaliza entre una escala entre cero y uno utilizando la siguiente formula

(2)

En esta etapa, el asistente inteligente es donde hará más énfasis debido a la amplia gama de fases en las que puede estar involucrado el usuario, cada uno de los métodos y fases anteriormente descritos, estarán disponibles y serán recomendados de acuerdo a la retroalimentación del usuario dada durante las primeras 2 etapas.

#### Reducción de dimensionalidad y selección de características

Reducir la dimensionalidad del conjunto de datos a analizar, ya sea la cantidad de atributos que tienen los registro o el número de registros que tienen en sí, permite reducir costo computacional al momento de aplicar algoritmos de minería de datos sobre dicho conjunto de datos (Chizi & Maimon, 2009, pág. 83). Teóricamente, conocer la relación funcional exacta entre el coste y el error puede señalar el clasificador ideal (es decir, un clasificador que produzca una tasa de error mínima ε∗ y costes h∗ a derivar (Chizi & Maimon, 2009, pág. 83).

Por otro lado, la selección de características es un problema muy relacionado con la reducción de la dimensionalidad, el objetivo de este es identificar atributos relevantes para el análisis y descartar los que son redundantes o no son relevantes, lo que conduce a algoritmos de minería más eficientes (Chizi & Maimon, 2009, pág. 84)*.*

Puede haber 4 razones por el cual en un proceso de KDD se reduce la dimensionalidad de un conjunto de datos y se filtran atributos: disminuir el coste de aprendizaje del modelo de minería de datos a utilizar; Incrementar el rendimiento del modelo de minería de datos a utilizar; reducir las dimensiones irrelevantes del conjunto de datos a analizar o reducir las redundancias del conjunto de datos a analizar (Chizi & Maimon, 2009, págs. 84-85)

Entre los métodos para la selección de características se encuentran

* Método FOCUS
* Método LVF
* Filtrar de características a través de discretización
* El uso de algoritmos de aprendizaje sobre otros algoritmos de aprendizaje para la selección de características
* Método RELIEF
* Métodos de envoltura, ya sea a través de árboles de aprendizaje, basados en instancias o en clasificadores de bayes

El asistente inteligente utilizará métodos de envoltura para filtrar características

### Algoritmos de minería de datos y análisis de datos

A pesar de que el asistente inteligente llegará hasta la fase de pre procesamiento de datos en cuanto a recomendaciones, es necesario conocer la etapa de minería de datos y los algoritmos relacionados a esta debido a que el pre procesamiento tiene objetivo final un mejor rendimiento de estos algoritmos.

En primer lugar debemos diferenciar entre lo que es una tarea de minería de datos y un método o algoritmo de minería de datos, ejemplo “Clasificar los diferentes reportes entregados por el cuerpo administrativo de la universidad católica Andrés Bello de acuerdo a una escala de relevante, óptimo, sin formación”, esto sería una tarea de clasificación y se podría resolver mediantes métodos de árboles de decisión, redes neuronales, entre otros (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 139).

A continuación se nombrarán todas las tareas que forman parte de la minería de datos y las vertientes dentro de ellas

#### Tareas predictivas

Las tareas predictivas de acuerdo a Orallo J., Ramírez Quintana M. J & Ferri Ramírez C. (2004) “se trata de problemas y tareas en los que hay que predecir uno más valores para uno o más ejemplos. Los ejemplos en la evidencia van acompañados de una salida (clase, categoría o valor numérico) o un orden entre ellos” (pág. 139). Las tareas predictivas se pueden definir de manera más concreta de acuerdo a varían de acuerdo a como sean presentados los datos y los valores de salida (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 139). Entre las que se encuentran

##### Clasificación

Las tareas de clasificación tiene como objetivo, dado un conjunto (donde “e” es un elemento de entrada perteneciente a “E”, que son todos los valores de entrada y “s” es el elemento de salida correspondiente de salida que pertenece a “S” el son todos los posibles valores resultantes), entrenar una función λ: E → S denominada clasificador, que permitirá, una vez entrenada, asignarle a cualquier valor de entrada no etiquetado un valor sin salida (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 139).

Un ejemplo de un uso recurrente de este tipo de tarea es la clasificación de mensajes de *spam* (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 139).

##### Categorización

Las tareas de categorización buscan asignar correspondencia a cada uno de los elementos dada una función , la diferencia principal entre categorización y clasificación es que tarea permite asignarle distintos valores de salida a cada uno de los elementos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 140).

Un ejemplo de esto es cuando se tienen objetos que realizan diversas funciones, un objeto computadora por ejemplo, puede categorizarse de acuerdo a sus funcionalidades en, ofimática, entretenimiento y comunicación.

#### Tareas descriptivas

Las tareas descriptivas, son aquellas donde los datos son presentados sin ninguna etiqueta, es decir presentan como un conjunto , el objetivo por lo tanto no es predecir sino describir el conjunto presentado, entre las tareas descriptivas se tienen

##### Agrupamiento (también conocido como clustering)

De acuerdo a Orallo, Ramírez Quintana M. & Ferri Ramírez. (2004) “el objetivo de esta tarea es obtener grupos o conjuntos entre los elementos de δ, de tal manera que los elementos asignados al mismo grupo sean similares” (pág.142). Similar a la función de una tarea de clasificación, esta función se define como λ: E → S, con la diferencia de que el conjunto S se crean o se inventan (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 142).

##### Reglas de asociación

Las reglas de asociación es una tarea similar a las correlaciones y factorizaciones sólo que para los atributos nominales, las reglas de asociaciones se definen de la siguiente forma: si, “ … ^ entonces “donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 143).

De acuerdo a Orallo, Ramírez Quintana, & Ferri Ramírez (2004)

*“*Existen muchas variantes de las reglas de asociación: reglas de asociación negativas (se pueden incluir desigualdades, lo que tiene interés cuando los atributos pueden tener más de dos valores posibles), reglas de asociación secuenciales (cuando las asociaciones no ocurren en el mismo momento sino en sucesivos registros en un intervalo de tiempo), multinivel (se consideran las categorías de los productos, por ejemplo), etc.” (pág. 143)

Los métodos que corresponden estas tareas, a pesar de que algunos forman parte de más de dos grupos debido a su flexibilidad, se pueden clasificar de manera muy general en diversos grupos entre los que se encuentran:

#### Métodos basados en frecuencias y tablas de contingencia

De acuerdo a Orallo J., Ramírez Quintana M. J & Ferri Ramírez C. (2004) “se basan en contar la frecuencia en la que dos o más sucesos se presenten conjuntamente” (pág. 146). Este tipo de métodos utilizan algoritmos que recogen un conjunto de datos y los cuentan a partir de cierto umbral, un ejemplo de esto es el algoritmo “a priori” (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 146).

#### Métodos basados en árboles de decisión y sistema de aprendizaje con reglas

Estos métodos se basan en 2 clases de algoritmos, los algoritmos denominados “divide y vencerás” (como el algoritmo de aprendizaje basado en árboles de decisiones y los algoritmos denominados “separa y vencerás” (como el CN2) (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 146).

#### Métodos basados en casos, en densidad o distancia

De acuerdo a Orallo J., Ramírez Quintana M. J y Ferri Ramírez C. (2004) “son métodos que se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos (los casos más similares), de una manera más sofisticada, mediante la estimación de funciones de densidad” (pág. 147), entre los métodos de este grupo más populares se encuentra el K medias

K medias de acuerdo a Orallo, Ramírez Quintana, & Ferri Ramírez (2004) “se trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar, sin etiquetar” (pág. 432), La idea del K medias es situar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares (Moody & Darken, 1989, págs. 281-294).

Una vez que todos los prototipos han sido ubicados, todos los nuevos datos son comparados son estos y asociados al más próximo en los términos de una distancia elegida por el usuario, generalmente una distancia euclídea. (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 432).

De acuerdo a Orallo, Ramírez Quintana, & Ferri Ramírez (2004) “las regiones se definen minimizando la suma de las distancias cuadráticas entre cada vector de entrada y el centro de su correspondiente clase, representado por el prototipo correspondiente” (pág. 432)

El algoritmo selecciona arbitrariamente particionales iniciales de forma que cada conglomerado o cluster tenga al menos 1 ejemplo, los centros de cada uno de estos conglomerados se calculan como la media de los datos de esos conglomerados y va cambiando a medida que el algoritmo se va ejecutando e introduciendo nuevos datos (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 432).

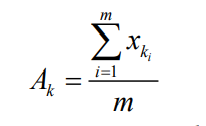
El procedimiento de este algoritmo de acuerdo a Orallo, Ramírez Quintana, & Ferri Ramírez (2004) es el siguiente:

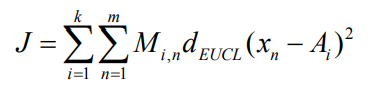
Primeramente se calcula, para cada ejemplo , el prototipo más próximo y se incluye en la lista de ejemplos de dicho prototipo

 (3)

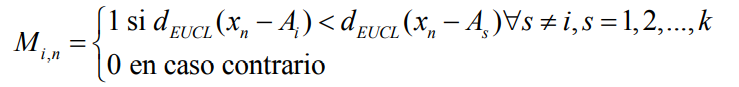
Después de haber introducido todos los ejemplos, cada prototipo tendrá un conjunto de ejemplos a los que representa (4)

Seguidamente se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos

 (5)

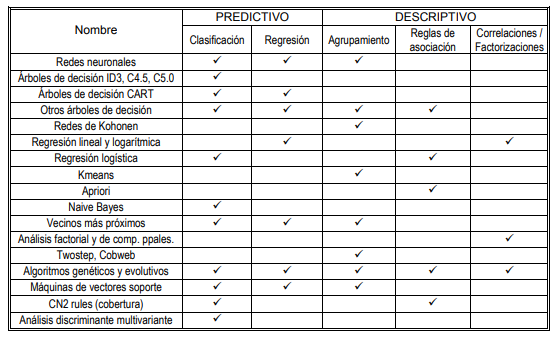
Y por último se repite el procedimiento hasta que ya no se desplazan los prototipos. Mediante este algoritmo el espacio de datos de entrada se divide en K clases o regiones y el prototipo de cada clase estará en el centro de la misma (págs. 432-433). Los centros se calculan buscando minimizar las distancias cuadráticas euclídeas entre los datos de entrada y el centro más cercano, es decir maximizando el valor J:  (6)

Donde es el conjunto de patrones, es la distancia euclídea, es el ejemplo de entrada , es el prototipo de la clase , y es la función de pertenencia del ejemplo a la región de forma que vale 1 si el prototipo es el más cercano al ejemplo y 0 en caso contrario, es decir

 (7)

Como se nombró anteriormente estos métodos pueden ser aplicados para una o más tareas como se muestra en la *tabla 2*

Tabla.2   
Comparativa de métodos con respecto a las tareas que pueden ser utilizados



Tomado de *Introducción a la minería de datos*, (pg. 148), Orallo J., Ramírez Quintana M. J y Ferri Ramírez C (2008).

### Visualización e interpretación de los resultados obtenidos

Comprender los resultados obtenidos a través de los algoritmos de minería y análisis de datos es una tarea fundamental (Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 523).

### Sistemas de soporte inteligentes

Una vez teniendo una idea clara acerca de las fases y etapas del descubrimiento de conocimiento donde el asistente inteligente se verá involucrado, se definirá el funcionamiento del mismo, las ventajas de tener uno sobre el proceso de descubrimiento de conocimiento y el método de inteligencia que se utilizará para el desarrollo del mismo.

La integración del proceso de descubrimiento de conocimiento con los sistemas de soporte de decisiones reducen la incertidumbre en la toma de decisiones durante el proceso de análisis (Chen, 2004, pág. 104), adicionalmente llevan al incremento del rendimiento de los sistemas de toma de decisiones (Mladenic, Lavrač, Bohanec, & Moyle, 2002, págs. 225-230), así como mejorar y ayudar a crear distintos enfoques que se tienen de los problemas actuales en donde se utilizan este tipo de sistemas, permitiendo fusionar el conocimiento de los expertos con el conocimiento extraído de los datos.

De acuerdo a (Philliips-Wren, 2013, págs. 26-28) los sistemas que utilizan inteligencia artificial para el apoyo de la toma de decisiones son llamados sistemas de soporte inteligente o *intelligent decision support system* en inglés (IDSS), de acuerdo a Simon (1977) este tipo de sistemas se basan en el modelo de 4 fases: 1) Inteligencia, 2) Diseño, 3) Elección, 4) Implementación.

Durante la fase de inteligencia, el tomador de decisiones desarrolla una comprensión del problema, articula la decisión a tomar y recopila información relevante para el problema de decisión (Philliips-Wren, 2013, pág. 28).

La fase de diseño se caracteriza por identificar variables importantes para el problema de decisión, definir criterios para la decisión, especificar relaciones entre variables, desarrollar un modelo de decisión que se puede utilizar para evaluar alternativas y explorar decisiones alternativas (Philliips-Wren, 2013, pág. 28).

En la fase de elección, el tomador de decisiones evalúa alternativas y selecciona una decisión que mejor cumpla con los criterios de decisión. La fase final de implementación (a veces llamada Revisión) es la fase en la que el tomador de decisiones considera las consecuencias de la decisión, desarrolla un plan de implementación, asegura recursos y actúa sobre la decisión (Philliips-Wren, 2013, pág. 28).

Las fases proceden generalmente de forma secuencial, aunque hay ciclos de retroalimentación entre las fases de modo que el tomador de decisiones puede volver a una fase anterior o comenzar otro proceso de decisión basado en resultados de la implementación (Philliips-Wren, 2013, pág. 28).

Los sistemas de soporte de decisiones tienen distintos nombre y varían para llegar al objetivo y el público para el cual fueron diseñados, entre estos nombres están: sistemas expertos, sistemas de soporte ejecutivo, sistemas de soporte de decisiones inteligente, sistemas de soporte de decisiones adaptivo, entre otros (Philliips-Wren, 2013, pág. 29). La inteligencia artificial a menudo es usada para expandir las capacidades de estos con la ayuda de diversos enfoques como lo son el de lógica difusa, o el de minería de datos y aprendizaje estadístico (Philliips-Wren, 2013, págs. 30-31). A continuación de describe el enfoque de minería de datos y aprendizaje estadístico haciendo énfasis en el método que se utilizará la elaboración de este trabajo especial de grado.

#### Minería de datos y aprendizaje estadístico

El enfoque de la minería de datos y aprendizaje estadístico se basa en analizar una gran cantidad de datos observables para encontrar relaciones no triviales y luego resumir estos datos de forma que sea entendible y útil para el dueño o dueños de los datos (Zopounidis & Doumpos, 2013, pág. 9). Este enfoque a su vez posee un método llamado basado en reglas y árboles de decisiones:

#### Basados en reglas y árboles de decisiones

Este método utiliza modelos basados en reglas y árboles de decisiones, Cada regla de decisión “Si” y “sino” se compone de una parte de condición que especifica un perfil parcial sobre un subconjunto de criterios con los que se compara una alternativa utilizando la relación de dominio, y una parte de conclusión que sugiere una recomendación de decisión (Zopounidis & Doumpos, 2013, pág. 11).

Este será el método utilizado para la elaboración de la inteligencia del asistente debido a que permite al usuario entregar retroalimentación constante durante el proceso, lo que permitirá que el asistente pueda sugerir fases de forma iterativa, logrando que el usuario finalmente que el usuario pueda sacar el máximo rendimiento del algoritmo de minería de datos por el cual sus datos pasarán.

Una vez definidos los métodos a utilizar para la elaboración del asistente inteligente se describen las herramientas para la construcción del mismo

#### Encadenamiento hacia adelante

El encadenamiento hacia adelante es una estrategia para la creación de reglas que consiste en recabar la mayor cantidad de datos acerca del problema para llegar a una decisión (Durkin, 1994, págs. 100-101).

## Terminología básica

### Vista minable

La vista minable es la tabla que contiene todos los datos necesarios que son o serán analizados

### Experiencia de desarrollador

Se define como experiencia de desarrollador, la facilidad que tiene el desarrollador para adaptarse y manipular las herramientas sobre las cuales está trabajando (Fagerholm & Münch, 2012)

# Capitulo III

# Marco Metodológico

## Tipo de investigación

El presente trabajo de grado propone el desarrollo de un asistente inteligente (un sistema que tomando en cuenta los datos y el problema a resolver recomiende la decisión más óptima) que guíe al usuario novato en la toma de decisiones durante un proceso de análisis de datos.

Este tipo de investigación se califica como proyectiva, debido a que estas buscan la elaboración de una propuesta, un programa, un procedimiento, entre otros como una solución a un problema planteado (Barrera de Hurtado, 2000, pág. 567). Este tipo de investigación tiene como característica abordar problemas prácticos y resolverlos mediante procesos de investigación (Barrera de Hurtado, 2000, pág. 567)

## Técnicas de recolección de datos

Una técnica de recolección de datos es un procedimiento o forma particular de obtención de datos o información que sean de utilidad en el desarrollo de la investigación, para este trabajo, durante su desarrollo se hará uso de entrevistas informales a expertos del área, revisión documental, lluvia de ideas y el uso de prototipos para la obtención de requerimientos (Fidias, 2012, págs. 72-74).

## Metodología de desarrollo

Para la realización de este trabajo de grado, se hace uso de una metodología basada en el modelo prototipo, con un enfoque de desarrollo evolutivo. El modelo de prototipo de acuerdo a García Peñalvo F. & García Holgado A. (2008) “es un modelo experimental de un componente de un sistema que tiene los suficientes elementos que permiten su uso” (pág. 17) y es ideal para probar suposiciones de características que deben ir en el sistema, evolucionan a través de un proceso iterativo y son brinda flexibilidad a la hora de introducir requerimientos (García Peñalvo & García Holgado, 2018, pág. 23).

De acuerdo a (Laudon & Laudon, 2016, pág. 507) el prototipo consiste en:

1. Identificar los requerimientos del usuario: El diseñador del sistema (por lo general, un especialista en sistemas de información) labora con el cliente únicamente lo suficiente para capturar las necesidades primordiales de información del usuario.
2. Desarrollar un prototipo inicial: El diseñador del sistema construye de la forma más rápida posible un producto mínimo viable funcional
3. Usar el prototipo: Se anima al usuario a trabajar con el sistema para determinar qué tan bien cumple el prototipo con los requerimientos especificados en el paso 1 y proveer retroalimentación constante al respecto.
4. Revisar y mejorar el prototipo: El creador del sistema anota todos los cambios que el usuario solicita, y refina apropiadamente el prototipo. Después de que se ha revisado el prototipo, el ciclo regresa al paso 3. Los pasos 3 y 4 se repiten hasta que el usuario queda satisfecho.

La metodología basada en prototipo se adecúa al desarrollo de este trabajo de grado debido a que permite la retroalimentación de los usuarios a los que va dirigido mientras se desarrolla el mismo, lo que permite una mejor adaptabilidad del lenguaje utilizado en el asistente, además, permite identificar y priorizar partes del proceso de descubrimiento de conocimiento en donde los usuarios tienen más dudas.

## Procedimiento metodológico

Para el desarrollo de este trabajo se procedió a hacer de forma secuencial los objetivos específicos descritos a continuación:

### Revisar el estado del arte de la construcción de asistentes inteligentes y de análisis de datos

Para la revisión del estado del arte se utiliza la técnica de revisión documental, la cual “el proceso basado en la búsqueda, recuperación, análisis, crítica e interpretación de datos secundarios, es decir, los obtenidos y registrados por otros investigadores en fuentes documentales: impresas, audiovisuales o electrónicas” (Fidias, 2012, pág. 27). Esta técnica se utilizara aplicando la metodología de revisión del estado del arte la cual consiste en los siguientes cuatro pasos

* Contextualización
* Clasificación
* Categorización
* Análisis

La contextualización consiste en la definición de criterios acerca del material a revisar (Molina, 2005, pág. 74).

La clasificación consiste en, teniendo en cuenta los criterios definidos previamente, el tipo de documentos a realizar revisar, las líneas de investigación a tomar en cuenta y las disciplinas de interés para la resolución del problema planteado (Molina, 2005, pág. 74). Para esto se toman en cuenta libros, artículos científicos y material en línea que cumplan con los criterios anteriormente descritos que tengan como alcance la implementación y el desarrollo de asistentes inteligentes, procesos de análisis de datos y descubrimiento de conocimiento.

La categorización consiste en la recuperación de la información relevante (Molina, 2005, pág. 74) que permita resolver el problema dentro de los libros, artículos científicos y material en línea seleccionados en la etapa de categorización.

Mientras que el análisis consiste en la transferencia del conocimiento adquirido a dar con la solución de la problemática planteada en este trabajo.

Al finalizar este objetivo, se obtiene un resumen de como realizar un proceso de análisis de datos y los pasos de forma detallada hasta el alcance del asistente.

### **Definir un marco de trabajo para el análisis de datos basándose en el estado del arte previamente revisado**

Mediante la revisión documental acerca de metodologías para definir un marco de trabajo para el análisis de datos no se encontró documentación al respecto, por ello considerando el factor tiempo, se utiliza una metodología heurística.

La metodología heurística es un conjunto de pasos que se utilizan para buscar y resolver problemas cuando es difícil encontrar información precisa o satisfactoria , estos pasos se definen de la siguiente manera: El primer paso es la comprensión del problema, definir la incógnita y a resolver y las condiciones para definir dicha incógnita, el segundo paso es definición de un plan para resolver la incógnita teniendo en cuenta los datos encontrados, el tercer paso es la ejecución del plan y el cuarto paso es la examinación de la solución obtenida (Polya, 1989, pág. 17).

Al finalizar este objetivo, se obtiene un paso detallado acerca de los métodos, algoritmos y pasos que deben ser aplicados en el proceso de la construcción de un asistente inteligente así como la integración del conocimiento obtenido en la revisión del estado del arte al asistente inteligente.

### Construir un asistente inteligente tomando como base la definición propuesta

Teniendo el marco de trabajo definido se diseña el asistente inteligente sobre los métodos, herramientas y procesos descritos previamente dentro del marco teórico.

Al culminar de esta etapa se obtiene un asistente inteligente que es capaz de guiar al usuario sobre procesos de análisis de datos bajo un marco de trabajo previamente definido.

### **Validar el marco de trabajo implementado en el asistente inteligente mediante un caso de estudio**

Se validará el funcionamiento del asistente mediante un caso de validación de caja negra (Suen, Grogono, Shingal, & Coallier, 1990, págs. 93-102) es decir, se hará una comparativa de la mejora mediante el pre procesamiento de los resultados de los algoritmos comprendidos y los datos extraídos sin pre procesamiento.

Para el modelos de agrupación se toma como medida de rendimiento la “suma de errores cuadrados dentro del grupo”, la cual es una medida que valida la calidad del agrupamiento cuanto menor sea (Expósito Izquierdo, Expósito Marquez, López Plata, Melián Batista, & Moreno Vega, s.f.) , además, se evaluara la duración del algoritmo en segundos para comprobar la mejora de su rendimiento.

Para el modelo de asociación se toma en cuenta la confianza de las reglas generadas es decir la probabilidad condicional en que esa regla se aplique, además de la frecuencia con las cuales se cumple esa regla dentro de los datos analizados (López Martínez, 2012)

# Capitulo IV

# Desarrollo y resultados

En el presente capítulo presentan los resultados alcanzados durante el desarrollo de los objetivos planteados, para desarrollar el asistente inteligente para soportar la toma de decisiones durante un proceso de análisis de datos. Para este fin, siguiendo la metodología definida para cada objetivo propuesto, se explican los resultados de las técnicas aplicadas y el análisis obtenido mediante recursos como ejemplos, tablas y gráficas. .

## Revisión del estado del arte

Estos criterios son los siguientes

* Relación al objetivo general de esta investigación: Material relacionado al área del análisis de datos, asistentes inteligentes y descubrimiento de conocimiento.
* Material hecho por uno o más autores avalados por una casa editorial, un centro de estudio u organizaciones reconocidas a nivel internacional relacionadas al análisis de datos o asistentes inteligentes.

**\*HASTA AQUÍ LO QUE ESCRIBES TAL COMO LO REDACTAS PARECE PARTE DE LA METODOLOGÍA… ES DECIR NO ESTÁN REDACTADAS COMO RESULTADOS NI COMO ANÁLISIS… YO TE RECOMIENDO SUBIR ESTO AL MARCO METODOLÓGICO Y AQUÍ REDACTAR ESTO COMO PASO A PASO EJECUTADO Y RESULTADOS ALCANZADOS, LUEGO CIERRAS CON UNOS PÁRRAFOS DONDE CONCLUYES**

**Esquema:**

* **subtitulo “revisar el estado del arte de la construcción de asistentes inteligentes”**
* **Subtitulo “revisar el estado del arte del análisis de datos\***

**/A PARTIR DE AQUÍ, ES PARTE DE TUS RESULTADOS… ANTES DEBES REFLEJAR CÓMO LLEGAS A ESTE RESULTADO, PASO A PASO DE LA METODOLOGÍA DESCRITA PARA ESTE OBJETIVO… LUEGO ANALIZAS Y EXPONES TUS CONCLUSIONES… /**

Teniendo en cuenta los resultados de este objetivo los cuales están redactados a lo largo del capítulo II se llegó a la metodología “CRIPS-DM” la cual es una metodología para la implementación del proceso de descubrimiento de conocimiento avalada por la comunidad europea que garantiza interoperabilidad en las industrias, ser de fácil acceso para los usuarios novatos e independiente de las herramientas utilizadas para el análisis de datos (Léon Guzmán, s.f.).

A continuación se expone el conocimiento relevante extraído de la línea de investigación del proceso de descubrimiento de conocimiento y análisis de datos tomando en cuenta como un asistente inteligente podría estar involucrado en el mismo, tomando como le referencia principal el documento “*CRISP-DM Step-by-step guide*”

1. Criterios de negocio: Antes de iniciar cualquier proceso de análisis de datos es importante tener que el asistente guíe al usuario mediante los siguientes puntos:
   1. Determinar los objetivos de negocio (contexto inicial, objetivos y criterios de éxito del negocio)
   2. Evaluar la situación (inventario de recursos, requerimientos supuestos, terminologías propias del negocio o el ámbito en el cual este se desarrolla, factores externos, entre otros)
   3. Establecer los objetivos de la minería de datos (objetivos y criterios de éxito)
   4. Generar el plan de proyecto (equipo involucrado experto en los datos a analizar, técnicas que se desean aplicar sobre los datos y fuentes de los mismos)
2. Comprensión de los datos: Para iniciar el proceso de análisis de datos el asistente debe tomar en cuenta que debe guiar el usuario a través de:
   1. La recopilación de las fuentes de las que se busca analizar los datos
   2. La descripción de los datos de las fuentes seleccionadas mediante técnicas como resumen y análisis de correlación.
   3. La exploración de los datos mediante técnicas como histogramas.
   4. La verificación de la calidad de los datos.
3. Técnica de minería de datos: A pesar de que el asistente se centra hasta la fase del pre procesamiento de los datos es necesario que el asistente pueda recomendar la técnica que se va a usar tomando en cuenta los objetivos del negocio y los objetivos de la minería de datos. Ya que a partir de la técnica seleccionada el asistente podrá recomendar métodos para lograr generar una vista minable óptima para la técnica seleccionada.
4. Criterios para la vista minable y métodos a aplicar: Es necesario que una vez seleccionada la técnica a utilizar el asistente recomiende métodos sobre una herramienta específica y que guíe al usuario en su uso.

## Definición del marco de trabajo

1. Definir un marco de trabajo para el análisis de datos basándose en el estado del arte previamente revisado.
2. Construir un asistente inteligente tomando como base la definición propuesta.
3. Validar el marco de trabajo implementado en el asistente inteligente mediante un caso de estudio.

El asistente fue construido para ayudar en el entorno de exploración de datos WEKA debido a la amplia documentación del mismo, y su facilidad de uso

Para el marco de trabajo se decidió trabajar sobre el enfoque de asistentes inteligentes basados en reglas, de acuerdo a (Durkin, 1994, pág. 168) estos deben contener 3 módulos:

* Base de conocimiento: El cual es un modelado de la memoria a largo plazo del humano como un conjunto de reglas
* Memoria de trabajo: El cual es un modelado de la memoria a corto plazo del humano así como el conjunto de hechos sobre las características del problema del mismo los cuales son proveídos al asistente para inferir sobre los mismos y pueda disparar reglas o procedimientos que se deben realizar
* Motor de inferencia: El cual es un modelado del razonamiento humano combinando hechos del problema contenidos en la memoria de trabajo con reglas contenidas en la base de conocimiento para inferir información o procedimientos que se deben realizar.

Adicionalmente las reglas se construyeron mediante la estrategia de encadenamiento hacia adelante, esta estrategia consiste en recabar la mayor cantidad de datos posibles para llegar a una decisión (Durkin, 1994, págs. 100-101), esto permite al asistente guiar al usuario por cada una de las fases del proceso mientras va recaudando información que le permitan sugerir los siguientes pasos a tomar durante el proceso de análisis de datos.

Adicionalmente el asistente inteligente debe contener los siguientes módulos siendo el último en esta lista opcional:

* Interfaz de usuario: El medio por el cual el usuario interactúa con el asistente y este recauda información acerca de problema.
* Interface de desarrollador: El medio por el cual es provista el conocimiento para futuras iteraciones.
* Módulo de explicación: El subsistema responsable de explicarle al usuario acerca de la razón de la recomendación dada.

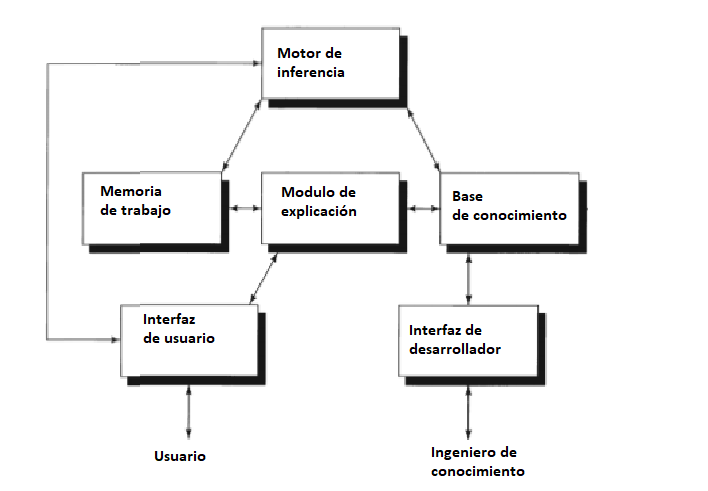
En la figura 3 se muestra un diagrama de cómo estos módulos deben estar interconectados, tomando en cuenta el ingeniero de conocimiento, el cual es el encargado de mantener el sistema experto.

Figura 3 Arquitectura de asistente inteligente. Adaptado de “Expert systems: Design and development” (1994) por Durkin J. (p. 170).

### Definición de criterios para los problemas de análisis de datos

A continuación se enumeran los criterios seleccionados por cada una de las fases del proceso de análisis de datos donde el asistente inteligente estará apoyará, tomando en cuenta la metodología CRIPS-DM 1.0 el cual es una metodología del procedimiento de KDD (Plotnikova, Dumas, & Milani, 2020)

#### Entendimiento del negocio

Según Chapman, y otros (2000, págs. 16-19 y 35-41) el entendimiento se divide en 4 tareas:

* Determinación de los objetivos del negocio: Descripción de los objetivos comerciales de la organización, esto le permitirá al analista y al asistente entender el porqué del análisis, generando así, tres entregables:
  + Organigrama organizacional: Descripción acerca de la estructura del negocio y las personas claves involucradas con los datos.
  + Problema del área: Análisis del problema en términos de negocio, así como la identificación del área de la organización al que está asociado así como la definición de expectativas y necesidades que este genera.
  + Solución actual: Análisis e identificación de la solución que tiene el negocio sobre este problema actualmente y ventajas y desventajas de la misma
  + Objetivos de negocio: Descripción del problema en términos de análisis de datos, especificación de objetivos a nivel de negocio relacionados al análisis de datos y beneficios esperados del análisis.
  + Criterios de negocio: Especificación de criterios de éxito del proyecto de análisis de datos en términos de negocio así como identificación de quien valida esos criterios de éxito dentro de la organización en la cual se realiza en proyecto de análisis de datos.
* Evaluación de la situación: Evaluación acerca de los recursos disponibles para el análisis de datos, recolección de fuentes de datos, evaluación de suposiciones sobre los datos y compilación de terminología utilizada en la organización, esto le permitirá al analista y al asistente entender sobre los datos los cuales analizará y la disponibilidad de los mismos, de esta tarea derivan cuatro entregables:
  + Inventario de recursos: Evaluación acerca del hardware a utilizar para el análisis de datos así como el utilizado en el negocio de forma diaria.
  + Fuentes de datos y conocimiento: Identificación de las fuentes de datos, conocimiento relevante y tipo de conocimiento dentro de la organización (expertos en los datos, documentación al respecto, bases de datos, entre otros).
  + Requerimientos, supuestos y limitantes: Identificación sobre suposiciones acerca de los datos, la calidad de los mismos y los factores externos que los afectan
  + Terminología: Compilación de glosario de términos relevantes para el proyecto de análisis de datos y de la organización
* Determinación de objetivos del proceso de análisis de datos: Especificación acerca los objetivos del análisis de datos, tareas de análisis de datos relacionadas a los objetivos de negocio previamente definidos y definición de criterios de éxito del análisis de datos, de esta tarea se derivan dos entregables:
  + Objetivos del proceso de análisis de datos.
  + Criterios de éxito del análisis de datos.

En el anexo 1 se muestran las tareas, entregables y actividades que se requieren para esta etapa de acuerdo a la guía de usuario provista por Chapman y otros (2000, págs. 35-42) mientras que en la anexo 2 se muestra la conversión de este conocimiento al modelo de reglas para el asistente

#### Exploración de los datos

Según Chapman, y otros (2000, págs. 20-22 y 43-46) la exploración de datos se divido en 4 tareas:

* Colección de datos iniciales: Es la unificación de todos los datos en una tabla atributo-valor así como la verificación de la calidad de los mismos, esta tarea se deriva en tres entregables:
  + Lista de toda los datos a utilizar, así como la importancia o no de cada uno de los datos
  + Integración de los datos en una tabla atributo-valor: Para la integración de los mismos, se utilizan lenguajes de consultas como SQL, , de estar disponible, herramientas OLAP.
  + Listado de la calidad de los datos y los criterios seguidos
* Descripcición de los datos: Es la examinación de la superficie de los datos adquiridos, esto deriva en un entregable
  + Reporte de descripción de datos: Es la identificación de los tipos de valores contenidos dentro de la tabla atributo-valor así como su taxonomía, su dependencia con otros atributos y las estadísticas básicas de cada uno de ellos.
* Exploración de los datos: Esta tarea aborda las preguntas de la minería de datos, que pueden ser tratadas mediante la consulta, la visualización y la elaboración de informes. Entre ellas se encuentran: la distribución de atributos clave, por ejemplo el atributo objetivo de una tarea de predicción relaciones entre pares o números pequeños de atributos; resultados de agregaciones sencillas; propiedades de subpoblaciones significativas y análisis estadísticos sencillos, esto deriva en un entregable:
  + Reporte de la exploración de datos: Es la descripción de los resultados de esta tarea, incluyendo las primeras conclusiones o hipótesis iniciales y su impacto en el resto del proyecto. Si es necesario,se incluyen gráficos y diagramas que indican las características de los datos o conduzcan a subconjuntos de datos interesantes para un examen más profundo.
* Verificación de la calidad de los datos: Es la examinación de la calidad de los datos, abordando cuestiones como: ¿son los datos completos (cubren todos los casos necesarios)? ¿Son correctos o contienen errores y, si los hay, cuán comunes son? ¿Faltan valores en los datos? En caso afirmativo, ¿cómo se representan, dónde se producen y qué frecuencia tienen? De esta tareas se deriva un entregable:
  + Reporte de la calidad de los datos: Es la enumeración de los resultados de la verificación de la calidad de los datos; si existen problemas de calidad, la enumeración las posibles soluciones.

En el anexo 3 se identifican las tareas, entregables y actividades que se requieren para esta etapa de acuerdo a lo expuesto anteriormente y en el anexo 4 se muestra la conversión de este conocimiento al modelo de reglas para el asistente teniendo en cuenta las herramientas disponibles en WEKA.

#### Pre procesamiento de datos

Para la fase de pre procesamiento de datos, debido a la alta cantidad de métodos disponibles, se delimitó a 2 técnicas, variando los criterios de acuerdo a los tipos de datos manejados por estas técnicas, a continuación se listan y se describen los criterios de cada una de las técnicas seleccionadas.

##### Agrupamiento mediante algoritmo de K medias

El método de K medias trabaja con datos numéricos y distancias entre los mismos, teniendo en cuenta esto, se tiene que los criterios más relevantes que deben ser evaluados para fase de pre procesamiento son

* Valores faltantes: Al ser las K medias un algoritmo que agrupa tomando en cuenta las distancias de los valores de los atributos, es importante remplazar o eliminar los valores faltantes para que el resultado del algoritmo no sea afectado.
* Eliminación de valores extremos: Los valores extremos o valores anómalos afectan como se agrupan los valores por lo tanto deben ser eliminados.
* Normalización de datos: La normalización de los datos es necesaria cuando se utilizan fuentes de datos que contienen escalas distintas, un ejemplo común es cuando se tiene, en una misma tabla, edad, comprendida entre los diez (10) y cincuenta (50) y a su vez se tiene ganancias, las cuales normalmente se registran con los miles y millones.
* Reducción de dimensionalidad de datos: La reducción de la dimensionalidad de los datos, se utiliza para descartar datos que contienen poca información además de generar un mejor rendimiento el algoritmo.

En el anexo 5 se muestran las tareas, entregables y actividades necesarias de acuerdo a lo expuesto anteriormente y en el anexo 6 la conversión de ese conocimiento a reglas teniendo en cuenta las herramientas disponibles en WEKA.

##### Asociación mediante algoritmo a priori

El método de algoritmo a priori trabaja solo con datos nominales o con clases, teniendo en cuenta esto, los criterios más relevantes que deben ser evaluados durante la fase de pre procesamiento son

* Eliminación o remplazo de valores anómalos: La eliminación de valores permitirá una mejor precisión del algoritmo utilizado.
* Discretización de los datos: La tarea de discretización como lo fue explicado en el marco teórico permite convertir atributos numéricos a nominales, permitiendo así al algoritmo trabajar con estos datos
* Reducción de dimensionalidad: La tarea de reducción de la dimensionalidad de los datos permite al algoritmo descartar atributos que no son relevantes para el proceso permitiendo así crear reglas que son más entendibles para el usuario final o el analista de datos.

En el anexo 7 se muestran las tareas, entregables y actividades necesarias de acuerdo a lo expuesto anteriormente y en el anexo 8 la conversión de ese conocimiento a reglas teniendo en cuenta las herramientas disponibles en WEKA.

#### Aplicación de algoritmo

Debido a que la aplicación del algoritmo sobre los datos ya se aplica durante la fase de exploración, las reglas del asistente para esta etapa vienen siendo una repetición del último paso de esta fase sin embargo, se muestra la extracción de las reglas aplicada al asistente para esta fase en particular en el anexo 9.

## Construcción de asistente inteligente tomando como base la definición propuesta

### Criterios para selección de motor de reglas

Para el proceso de construcción de selección del motor de reglas se tuvieron en cuenta los siguientes criterios:

#### El motor de reglas debe seguir la estrategia de encadenamiento hacia adelante

Debido a la naturaleza de un problema como lo es el análisis de datos, el motor debe ser capaz de recabar la mayor cantidad de información acerca del conocimiento que tiene el usuario acerca de los datos y el contexto en el que se desenvuelve para poder dar una recomendación acertada.

#### El motor de reglas debe ofrecer una buena experiencia para el desarrollador

Este es un requisito que debe ser tomado en cuenta debido a que las reglas de los asistentes inteligentes en todo momento deben poder ser expandibles con el paso de tiempo a medida que la organización o los expertos ganan conocimiento acerca del problema para el cual los asistentes fueron diseñados (Durkin, 1994, pág. 294).

Para la construcción del asistente inteligente, tomando en los criterios anteriormente mencionados se hicieron 3 iteraciones para la selección del motor de reglas a continuación se describen cada una de ellas

### Drools

Drools de acuerdo a su página web es una herramienta en el lenguaje de programación JAVA que permite la creación de reglas en su sistema de reglas de negocio mediante su mesa de trabajo integrada con el entorno de desarrollo Eclipse (Overview, s.f.), sin embargo debido a la documentación poco clara acerca de cómo poner integrar dicha mesa de trabajo a Eclipse y entorno poco amigable para el desarrollador debido a la gran cantidad de paquetes requeridos para su funcionamiento, se descartó.

### Json-rules-engine

Esta es una librería de código abierto hecha para NodeJS, que trabaja sobre reglas, hechos y eventos en formato JSON, por lo cual resultaba una opción ideal, ya que permitía a su vez el desarrollo de una interfaz limpia y simple mediante el navegador web, sin embargo, presentó problemas de rendimiento después de manejar más de 10 reglas y hechos a la vez, por lo cual se terminó descartando.

### Algoritmo de reglas basándose en json-rules-engine

Aunque la librería de json-rules-engine presentó problemas de rendimiento el uso de archivos JSON permitía de manera sencilla crear y editar las reglas por lo se decidió crear una estructura de datos de tipo diccionario, el cual posee identificadores y valores únicos relacionados a estos (Ramírez, 2015).

 Esta estructura permite almacenar información de forma simple y entendible para el usuarios, además de navegar a través de ella, en el caso del asistente permite a su vez ofrecer recomendaciones o próximos pasos tomando en cuenta la respuesta del usuario. A continuación se muestra en la figura 4 el esquema planteado como base de conocimiento y sistema de reglas

Figura 4 Esquema para creación de reglas

A continuación se explican a detalle cada uno de los atributos

* nombreDeLaRegla: La clave nombreDeLaRegla puede ser remplazado con el nombre de la regla o información que se desea saber acerca del problema, esto permite al algoritmo dirigirse a una regla especifica de acuerdo dada por el usuario y a su vez contiene las siguientes claves
  + pregunta: la clave pregunta es el texto que se muestra en pantalla y que el usuario debe responder de manera afirmativa o negativa
  + Atributos Si y No: Estos atributos se refieren a la respuesta dada por el usuario de acuerdo al atributo pregunta, a su vez puede contener 2 estructuras de forma intercambiable
    - Estructura de regla: La estructura de regla como se muestra en el valor “Si” contiene 2 claves
      * tipo: esta clave va con el valor “regla”, lo cual le permite al algoritmo sabe que esta es una regla
      * regla: es el nombre de la regla al cual este debe buscar para seguir con el proceso.
    - Estructura de recomendación: La estructura de recomendación contiene 2 atributos
      * Tipo: el cual va con el valor “recomendación” para indicarle al algoritmo que esta estructura pertenece a una recomendación
      * recomendaciones: la clave recomendaciones contiene un arreglo de esquemas de que contienen 2 atributos
        + texto: Son las indicaciones que se le dan al usuario
        + imagen: Es la ruta de la imagen para darle contexto al usuario acerca de que botón o herramienta se refiere el asistente de ser necesario
      * proximoPaso: Esta clave va debe ser llenada con la siguiente regla que se debe hacer una vez se hayan aplicado todas las recomendaciones

A continuación en la figura 5 se muestra el algoritmo en pseudocódigo para el uso de esta estructura de datos

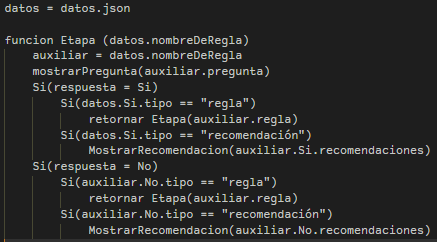


Figura 5 Algoritmo de asistente para cumplimiento de reglas utilizando el esquema planteado

Adicionalmente el asistente utiliza una función llamada siguientePaso la cual llamada a la función Etapa con la clave correspondiente a la clave “siguientePaso” de la recomendación que se está dando en ese determinado momento.

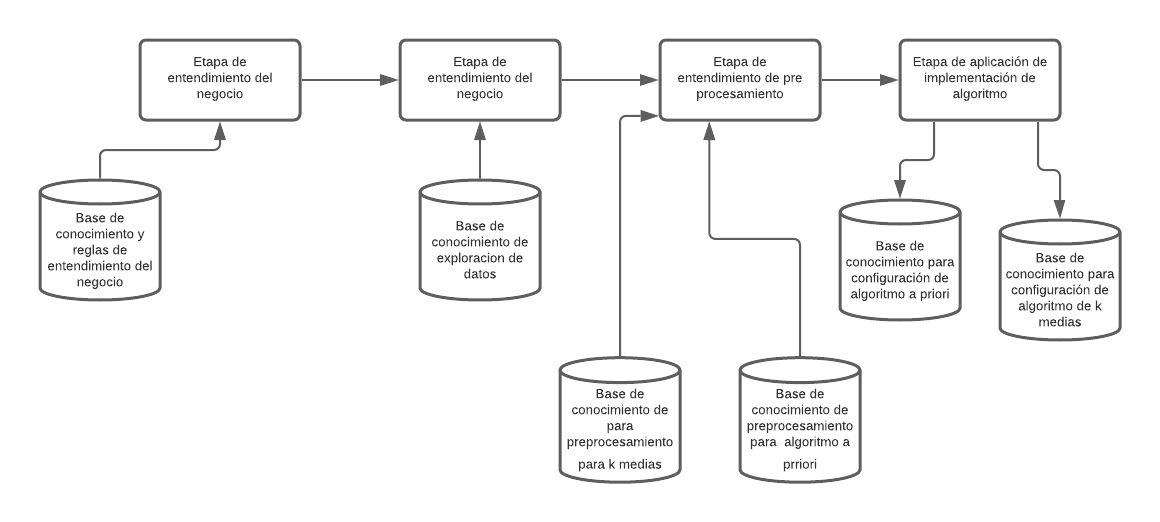
### Interfaz

La interfaz, como se mencionó en la definición de marco de trabajo, es una parte crucial del desarrollo de un asistente inteligente debido a que esta es la manera por la cual el asistente interactúa y recibe información por parte del usuario.

Para la creación de la interfaz se utilizó una interfaz web mediante el uso del marco de trabajo VueJS debido la popularidad del mismo y su facilidad de uso (Frontend frameworks , 2020).

### Arquitectura del asistente

La arquitectura utilizada para el asistente se presenta en la figura 6

Figura 6 *Esquema de arquitectura de asistente inteligente*

Donde cada etapa es un componente con una interfaz gráfica, asociada a una base de conocimiento en formato JSON.

## Validación de marco de trabajo implementado

Para la validación de estos casos de prueba se utilizaron los siguientes pasos con un sujeto de prueba ajeno al campo del análisis de datos:

* Paso 1: Etapa de obtención de requerimientos en el análisis de datos.
* Paso 2: Etapa de comprensión de los datos a analizar.
* Paso 3: Etapa de análisis de los datos.
* Paso 4: Etapa de evaluación de resultados de métodos utilizados utilizando las métricas de error de la suma de cuadrados para el algoritmo de k medias (el cual indica el nivel de variación total entre los componentes) y para el algoritmo a priori se utilizaron las métricas de nivel de confianza y frecuencia.

Estos pasos fueron realizados en primera instancia sin el asistente y luego el sujeto de prueba realizó el mismo análisis siguiendo las recomendaciones del asistente

### Primer caso de estudio sin asistente

#### Paso 1

El primer caso de validación consistió en un caso académico sobre los votos de los congresistas de la cámara de representantes de los Estados Unidos de América para el año 1986, este conjunto de datos incluyó los votos de cada uno de los congresistas sobre 16 tópicos distintos, y tenía como objetivo descubrir patrones de voto, esto de acuerdo a la número dos acerca de reglas de asociación con la herramienta WEKA de la universidad estadounidense Fordham (Universidad de Fordham, s.f.)

#### Paso 2

Los datos presentados en este conjunto de datos fueron nominales, por lo cual el sujeto de pruebas decidió que para extraer los patrones de voto utilizaría el algoritmo a priori debido a que este es un algoritmo adecuado para extraer patrones de los datos. En la figura 7 se muestran los datos cargados en el entorno WEKA así como los 17 atributos que corresponde este conjunto de datos.

#### Paso 3

En la etapa de análisis de datos el sujeto de prueba realizó el análisis mediante el algoritmo a priori, en la figura 8 se muestran los resultados obtenidos

#### Paso 4

En los resultados arrojados por el entorno WEKA se muestra que la confianza de estas reglas ronda entre 1 y 0.96 un porcentaje elevado de las mismas.

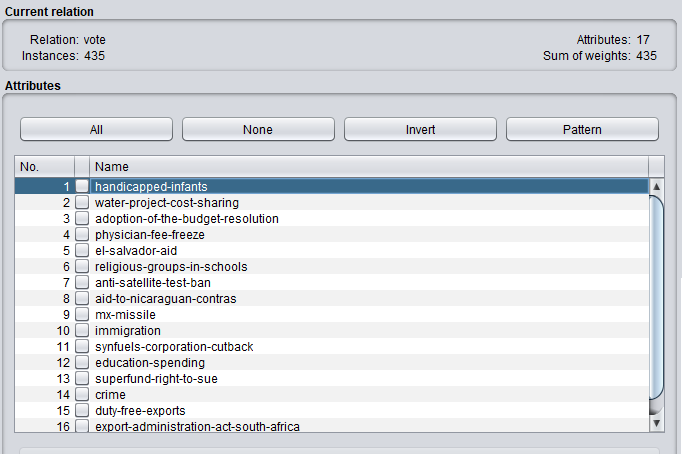


Figura 7 Muestra de datos académicos sobre los votos de los representantes del congreso de EE.UU para el año 1986

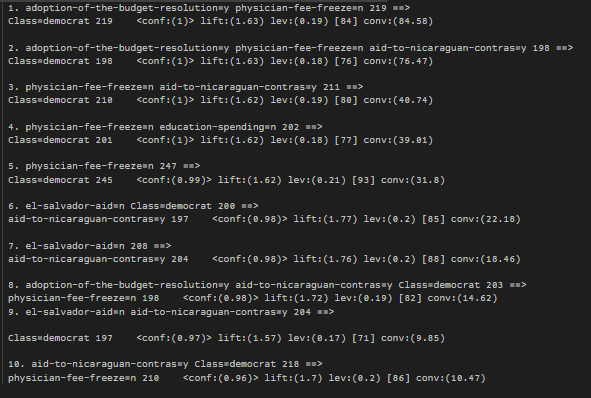


Figura 8 Resultados de algoritmo a priori sin asistente

### Primer caso de estudio con asistente

#### Paso 1

El paso 1, de validación se corresponde con la etapa 1 del asistente, en los anexos 10 y 11 se muestran preguntas hechas por el asistente durante el proceso, al finalizar esta etapa se extrajo la siguiente información:

Al tener un conjunto de datos académicos las preguntas relacionadas a la organización como la mostrada en el anexo 12 o el experto de datos no podían ser respondidas, la disponibilidad de los datos a analizar es del 100% del tiempo debido a que no afectaba a ningún sistema en producción, las fuentes de datos al ser de tipo académico no presentaron ningún término desconocido para el sujeto de pruebas y se centran en un solo conjunto de datos o una sola vista minable, la tarea de análisis de datos recomendada por el asistente fue la implementación de un algoritmo a priori como se muestra en el anexo 13.

#### Paso 2

Mediante las recomendaciones del asistente el usuario de prueba revisó la cantidad de atributos vacíos, como lo muestra la recomendación del anexo 14, verificó que los datos estuvieran integrados de manera correcta, estudió la taxonomía de los datos para verificar que estos fueran independientes unos de otros y además se aplicó el algoritmo para una revisión previa de los datos.

En el anexo 15 se muestra como el asistente guía al usuario para la aplicación del algoritmo obteniendo así los mismos resultados de la figura 8

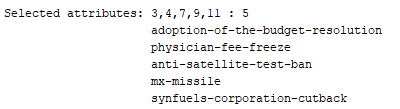
#### Paso 3

Para el análisis de datos, antes de proceder con el mismo el asistente sugirió realizar un pre procesamiento de los datos para corregir anomalías en los datos y mejorar la calidad de los mismos, adicionalmente como se muestra en el anexo 16, asistente recomendó una reducción de dimensionalidad, adicionalmente en la figura 9 se muestran los valores tomados por el algoritmo de selección de atributos recomendado por el asistente, por lo el sujeto de pruebas pasó a tener de 15 atributos a sólo 6.

Luego de esto el asistente recomendó utilizar nuevamente el algoritmo, como se muestra en el anexo 15, en la figura 10 se muestran los resultados obtenidos

#### Paso 4

Los resultados obtenidos de esta iteración le permitieron comprobar al sujeto de prueba que, la reducción de datos atributos no era necesaria, ya que, las reglas obtenidas después de la disminución de los atributos, en este caso en particular, disminuyeron la confianza de las reglas generadas, de oscilar entre los valores 1 y 0.96 en el análisis previo a tener una confianza entre 1 y 0.93.

*Figura 9* Atributos seleccionados por el clasificador

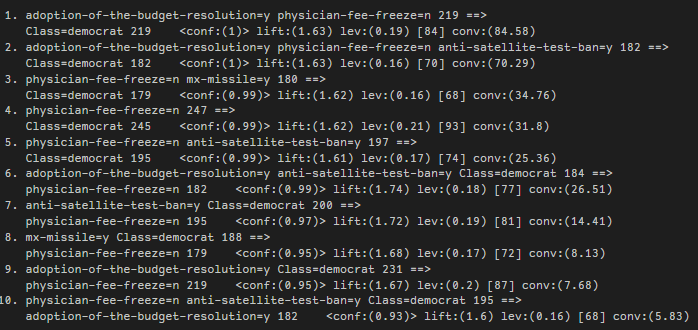


Figura 10 Reglas seleccionadas por el algoritmo a priori con ayuda del asistente

### Segundo caso de estudio sin asistente

#### Paso 1

El segundo caso de validación consistió en un caso académico de un conjunto de datos de vidrio que consta de 214 instancias y 10 atributos. Cada instancia representaba una pieza de vidrio, y su clase el tipo de vidrio. Con 7 tipos posibles, que corresponden a diferentes procesos de fabricación del vidrio. El objetivo de este análisis era agrupar el conjunto de datos para encontrar características comunes

#### Paso 2

En la etapa de conocimiento de datos, el sujeto de pruebas cargó sin problemas el conjunto de datos, se encontró con que la mayoría de los atributos eran de tipo numérico como se muestra en la figura 11

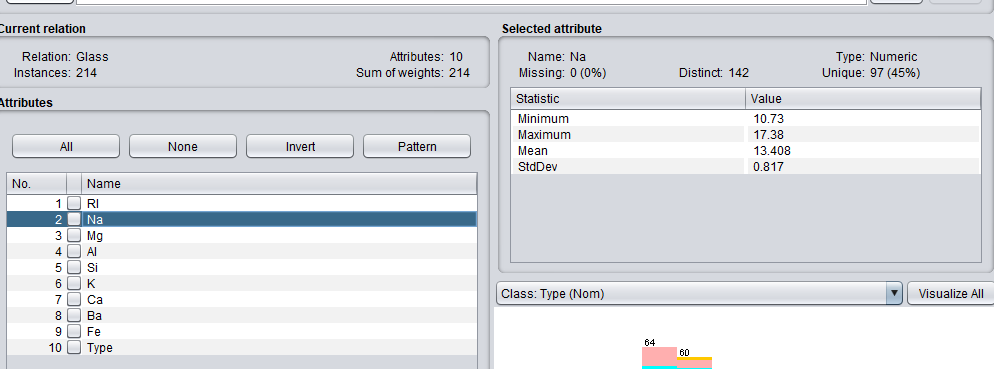


Figura 11 Muestra de datos de vidrios cargados al entorno WEKA

#### Paso 3

En la etapa de análisis de datos, el sujeto de pruebas realizó el análisis de los datos mediante el algoritmo de k medias, iterando múltiples veces sobre la cantidad de agrupamientos que debía definir, llegando a la conclusión de que la cantidad de agrupamientos debía ser 7 debido a cantidad de tipos de vidrios que había en el conjunto de datos. En la figura 12, se muestran los resultados obtenidos.

#### Paso 4

De acuerdo a los resultados obtenidos, sin ayuda del asistente, el sujeto de pruebas obtuvo una suma de error de los cuadrados de 49.94.

Este resultado indica que la variación de cada elemento del centro del conjunto al que pertenece, lo cual nos permite saber que aunque los resultados están agrupados poseen una variación alta entre ellos.

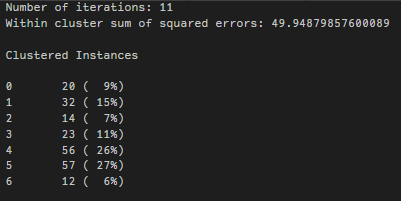


Figura 12 Resultados obtenidos de aplicación de algoritmo de K medias sin ayuda del asistente

### Segundo caso de estudio con asistente

#### Paso 1

El paso 1, de validación se corresponde con la etapa 1 del asistente, las preguntas hechas fueron las mismas que en el primer caso de validación con el asistente y la información extraída fue la siguiente:

Al tener un conjunto de datos académicos las preguntas relacionadas a la organización como la mostrada en el anexo 12 o el experto de datos no podían ser respondidas, la disponibilidad de los datos a analizar es del 100% del tiempo debido a que no afectaba a ningún sistema en producción, las fuentes de datos al ser de tipo académico no presentaron ningún término desconocido para el sujeto de pruebas y se centran en un solo conjunto de datos o una sola vista minable, la tarea de análisis de datos recomendada por el asistente fue la implementación de un algoritmo de agrupación como se muestra en el anexo 17 debido a que el objetivo del análisis traducido a objetivos de minería de datos era agrupar los datos, el sujeto de pruebas, llegó a esta conclusión después de haber paso por la pregunta mostrada en el anexo 18

#### Paso 2

Mediante las recomendaciones del asistente el sujeto de prueba revisó la cantidad de atributos vacíos del conjunto de datos, como lo muestra la recomendación del anexo 14, verificó que los datos estuvieran integrados de manera correcta, estudió la taxonomía de los datos para verificar que estos fueran independientes unos de otros y además se aplicó el algoritmo para una revisión previa de los datos.

En la se muestra como el asistente guía al usuario para la aplicación del algoritmo obteniendo así los resultados iguales a los de la figura 12.

#### Paso 3

Mediante las recomendaciones del asistente el sujeto de prueba redujo la cantidad de atributos de los, pasando de 10 atributos a 4 atributos, como se muestra en la figura 13, adicionalmente el usuario de prueba eliminó los valores extremos gracias a la recomendación del asistente, como se muestra en el anexo 19 y anexo 20, habiendo pasado así de 214 instancias a 164, adicionalmente, el sujeto de pruebas una vez habiendo hecho el pre procesamiento de los datos, aplicó nuevamente la tarea de k medias como se muestra en el anexo 21. En la figura 14 se muestran los resultados obtenidos

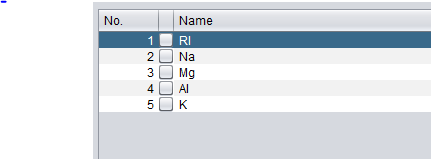


Figura 13 Caso estudio con atributos reducidos

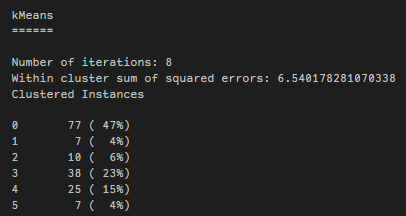


Figura 14 Resultados de aplicación de algoritmo de k medias después de pre procesamiento con ayuda del asistente

#### Paso 4

El usuario de pruebas pudo concluir que, tomando en cuenta los resultados posteriores al pre procesamiento mostraron una mejora significativa de 49.94 en la suma de cuadrados perfecto a 6.5, lo que indica que los valores están agrupados sobre los centroides de los grupos de una manera más uniforme

Como conclusión de los procesos de validación, se pudo demostrar que aunque no siempre se obtiene una mejora con los métodos recomendados por el asistente como lo fue en el primer caso de estudio, el asistente permitió que se tuviera una visión más amplia de los datos así como le permitió al usuario de pruebas aplicar filtros y métodos sobre los datos que sin ayuda del mismo, no hubiese tomado en cuenta como lo fue la detección y eliminación de valores anómalos.

# Capítulo V

# Conclusiones y Recomendaciones

## Conclusiones

Con el estudio y análisis del estado de arte se pudo estudiar el proceso de descubrimiento de conocimiento y la metodología CRIPS-DM los cuales fueron de vital importancia como punto de partida para el asistente inteligente para poder evaluar el tipo de preguntas y recomendaciones que debe hacer al usuario en el proceso de descubrimiento de conocimiento con el objetivo de mejorar la calidad de los datos.

La revisión documental acerca del funcionamiento de los asistentes inteligente y el enfoque de reglas y casos permitió crear un marco de trabajo que puede ser implementado en múltiples entornos más allá del proceso de descubrimiento de conocimiento.

La construcción del asistente inteligente permitió además evaluar las distintas alternativas que existen en el mercado y que pueden ser implementadas pero que sin embargo, no se adaptaban a los requerimientos de este trabajo especial de grado.

La validación mediante los casos de estudio permitió verificar la mejora de los de los resultados de las tareas de minería y como la aplicación del asistente sugirió otros puntos a tomar en cuenta durante el proceso de análisis de datos como lo fueron la verificación de la integración de los datos y la existencia de valores extremos que afectaba el análisis.

## Recomendaciones

Expandir la base la base de conocimiento del asistente en futuros trabajos, expandir las recomendaciones de los parámetros de los algoritmos de a priori y k medias así como incluir una mayor cantidad algoritmos a tomar en cuenta como otros algoritmos de clustering, otros algoritmos de selección de atributos y otro tipo de técnicas como las redes neuronales y algoritmos de regresión

Fomentar el estudio de los asistentes inteligentes combinando las áreas de psicología y afines para mejorar la redacción y términos utilizados en las recomendaciones con el fin de poder llegar a un público más amplio

Expandir el alcance del asistente inteligente a través de agentes, lo que permita una automatización en el proceso.

Expandir la las preguntas acerca de los datos involucrando asistente expertos en el área del análisis de datos y la exploración de conocimiento para poder dar recomendaciones más acertadas

# Bibliografía

Barrera de Hurtado, J. (2000). *Metodología de la investigación holística* (Tercera ed.). Caracas, Venezuela: Fundación Sypal.

Barret, V., & Lewis, L. (1978). *Outliers in statistical data.* John Wiley & Sons.

Berthold, M. &. (2007). *Intelligent data analysis: An introduction* (Segunda ed.). Berlin, Alemania: Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRIPS-DM 1.0 Step-by-step data mining guide.* USA: SPSS Inc.

Chen, K. (2004). Decision Support System for Tourism Development: System Dynamic Approach. *Journal of Computer Information Systems, 1*(45), 104-112.

Chizi, B., & Maimon, O. (2009). Dimension Reduction and Feature Selection. En O. Maimon, & L. Rokach, *Data Mining and Knowledge Discovery Handbook* (Segunda ed., págs. 83-100). Tel-Aviv: Springer.

Depool , R., & Monasterio, D. (2013). *Probabilidad y estadística Aplicaciones a la ingeniería.* Barquisimeto: UNEXPO.

Durkin, J. (1994). *Expert Systems: Design and development.* New York, USA: Macmillan Publishing Company.

Expósito Izquierdo, C., Expósito Marquez, A., López Plata, I., Melián Batista, B., & Moreno Vega, J. (s.f.). *Clustering basado en prototipos.* Obtenido de Universidad de la Laguna Aula, Campus Virtual: shorturl.at/yBSWY

Fagerholm, F., & Münch, J. (2012). Developer Experience: Concept and Definition. (págs. 73-77). Zurich: IEEE. doi:10.1109/ICSSP.2012.6225984

Fayyad , U., Smyth, P., & Piatetsky-Shapiro, G. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 3*(17), 82-88.

Fayyad, U., Piatetsky-Shapiro, G., & Uthurasamy, R. (2003). Data Mining: The Next 10 Years. *ACM SIGKDD Explorations Newsletter, 5*(2), 191-196.

Fayyad, U., Piatetsky-Shapiro, G., & Uthurasamy, R. (2013). Data Mining: The Next 10 Years. *ACM SIGKDD Explorations Newsletter, 5*(2), 191-196.

Fidias, G. A. (2012). *El proyecto de Investigación: Introducción a la medología científica* (Sexta ed.). Caracas: Editorial Episteme.

*Frontend frameworks* . (2020). Obtenido de State of of JS: https://2020.stateofjs.com/

Galhardas, H., Florescu, D., Sasha, D., Simon, E., & Saita, C.-A. (2001). *Declarative Data Clearning: Language Model, and Algorithms.*

García Peñalvo, F., & García Holgado, A. (21 de Febrero de 2018). *Repositorio de GRIAL: Metodologías de ingeniería de software.* Recuperado el 20 de Febrero de 2021, de Grial Repository: https://repositorio.grial.eu/handle/grial/1149

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter, 1*(1), 20-33. doi:https://doi.org/10.1145/846170.846172

Grzymala-Busse, J., & Grzymala-Busse, W. (2005). Handling Missing Attribute Values. En O. Maimon, & L. Rokach, *Data mining and knowledge discovery handbook* (Segunda ed.). Tel-Aviv, Estados Unidos: Springer.

Han, J., Kamber, J., & Pei , J. (2012). *Data mining Concepts and techniques* (Tercera ed.). Estados Unidos: Elsevier Inc.

Hinshelwood, S. (12 de Noviembre de 2008). *Business Partner Magazine*. Recuperado el 02 de 10 de 2020, de https://businesspartnermagazine.com/5-reasons-why-data-analysis-is-important-for-every-business/

Kantardzic, M. (2020). *DATA MINING Concepts, Models, Methods, and Algorithms* (Tercera ed.). New Jersey, Estados Unidos: Wiley.

Knight, K. (06 de Julio de 2015). *Report: Data scientists spend bulk of time cleaning up*. Recuperado el 15 de Octubre de 2020, de BizReport: http://www.bizreport.com/2015/07/report-data-scientists-spend-bulk-of-time-cleaning-up.html

Laudon, K., & Laudon, J. (2016). *Sistemas de información generencial* (Decimocuarta ed.). México: Person Educación.

Léon Guzmán, E. (s.f.). *Metodologías aplicadas al proceso de Minería de Datos.* Obtenido de Universidad Nacional de Colombia: https://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5\_Metodologias.pdf

López Martínez, A. (2012). *Análisis de patrones de navegación de los estudiantes dentro del Campus UOC.* (F. Xhafa, Ed.) Barcelona, España: Universitat Oberta de Catalunya. Obtenido de http://openaccess.uoc.edu/webapps/o2/bitstream/10609/13941/1/alopezmartinez\_TFC\_0612.pdf

Maletic, J. I., & Marcus, A. (2010). Data Cleansing: A Prelude to Knowledge Discovery. En O. Maimon, & L. Rokach, *Data Mining and Knowledge Discovery Handbook* (Segunda ed., págs. 19-32). Tel-Aviv: Springer.

Mladenic, D., Lavrač, N., Bohanec, M., & Moyle, S. (2002). *Data Mining and DecisionSupport: Integration and Collaboration.* Dordrecht: Kluwer Academic Publishers.

Molina, N. (2005). ¿Que es el estado del arte? *Ciencia y Tecnología de la Salud Visual y Ocular*, 73-75.

Moody, J., & Darken, C. (1989). *Fast learning in networks of locally tuned processing .* Neural Computation.

Orallo, J., Ramírez Quintana, M., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* Madrid, España: PEARSON EDUCACIÓN S.A.

Orr, K. (1998). Data Quality and Systems Theory. *Communications of the ACM, 41*(2), 66-71.

*Overview*. (s.f.). Recuperado el 10 de Abril de 2021, de Drools: https://www.drools.org/

Philliips-Wren, G. (2013). Intelligent Decision Support Systems. En D. Duompos, & E. Grigoroudis, *Multicriteria Decision Aid and Artificial Intelligence,* (págs. 25-44). Creta, Grecia: Wiley-Blackwell. doi:https://doi.org/10.1002/9781118522516.ch2

Plotnikova, V., Dumas, M., & Milani, F. (2020). *Adaptations of data mining methodologies: a systematic literature review.* PeerJ Computer Science. doi:10.7717/peerj-cs.267/supp-1

Puiu, T. (28 de Enero de 2020). *How Liverpool FC is using data science to dominate British Premier League*. Recuperado el 02 de 12 de 2002, de https://www.zmescience.com/science/how-liverpool-fc-is-using-data-science-to-dominate-british-premier-league

Ramírez, E. (2015). *Algoritmos y Estructura de datos.* Caracas, Venezuela: CCG-UCV.

Real Academia Española. (12 de Diciembre de 2020). *Dato | Definición | Diccionario de la Real Academia Española*, 23.4. Obtenido de Diccionario de la lengua española: https://dle.rae.es/dato

Resagratia. (20 de Junio de 2020). *Resagratia*. Recuperado el 2 de Diciembre de 2020, de https://resagratia.com/2020/06/the-importance-of-data-analysis/

Rupnik, R., Kukar, M., Bajec, M., & Krisper, M. (2006). DMDSS: Data mining based decision support system to integrate data mining and decision support. *28th International Conference on Information Technology Interfaces* (págs. 225–230). Cavtat: IEEE. doi:10.1109/ITI.2006.1708432

Serban, F., Vanschoren, J., Kietz, J.-U., & Bernstein, A. (Julio de 2013). A survey of intelligent assistants for data analysis. *ACM Computing Surveys, 45*(3), 2. doi:https://doi.org/10.1145/2480741.2480748

Srinivasan, S., Singh, J., & Kumar, V. (2011). Multi-agent based decision support system using data mining and case based reasoning. *International Journal of Computer Science Issues, 8*(4).

Suen, C., Grogono, P., Shingal, R., & Coallier, F. (1990). Verifying, validating, and measuring the performance of expert systems. *Expert Systems with Applications*, 93-102.

Timarán-Pereira, S. R., Hernández-Arteaga, S., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., & Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (págs. 63-86). Ediciones Universidad Cooperativa de Colombia. doi:http://dx.doi.org/10.16925/9789587600490

Universidad de Fordham. (s.f.). *Lab Exercise Two: Mining Association Rule with WEKA Explorer.* Bronx: Universidad Fordham. Recuperado el 15 de Mayo de 2021, de https://storm.cis.fordham.edu/~yli/documents/CISC4631Spring16/Weka\_LabTwo.pdf

Wielenga, D. (2007). Identifying and Overcoming Common Data Mining Mistakes. *SAS GLOBAL FORUM 2007*, (págs. 1-19). Cary.

Witten, I., Frank, E., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques.* Estados Unidos: Morgan Kaufmann.

Witten, I., Hall, M., Frank, E., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (Tercera ed.). Estados Unidos: Morgan Kaufmann.

Wright, E., & Daintith, J. (2008). *Artificial Intelligence - Oxford Reference*. Recuperado el 19 de Octubre de 2020, de Oxford Reference: https://www.oxfordreference.com/view/10.1093/acref/9780199234004.001.0001/acref-9780199234004-e-204

Zopounidis, C., & Doumpos, M. (2013). Computational intelligence techniques for multicriteria decision aiding: An overview. En C. Zopounidis, & M. Doumpos, *Multicriteria Decision Aid and Artificial Intelligence Links, Theory and Applications* (Primera ed., págs. 3-15). Creta, Grecia: John Wiley & Sons, Ltd.

# Anexos

**Anexo 1 Tareas, entregables y actividades de entendimiento de negocio**





**Anexo 2 Reglas basadas conocimiento adquirido de entendimiento de negocio**







**Anexo 3 Tareas, entregables y actividades de exploración de datos**



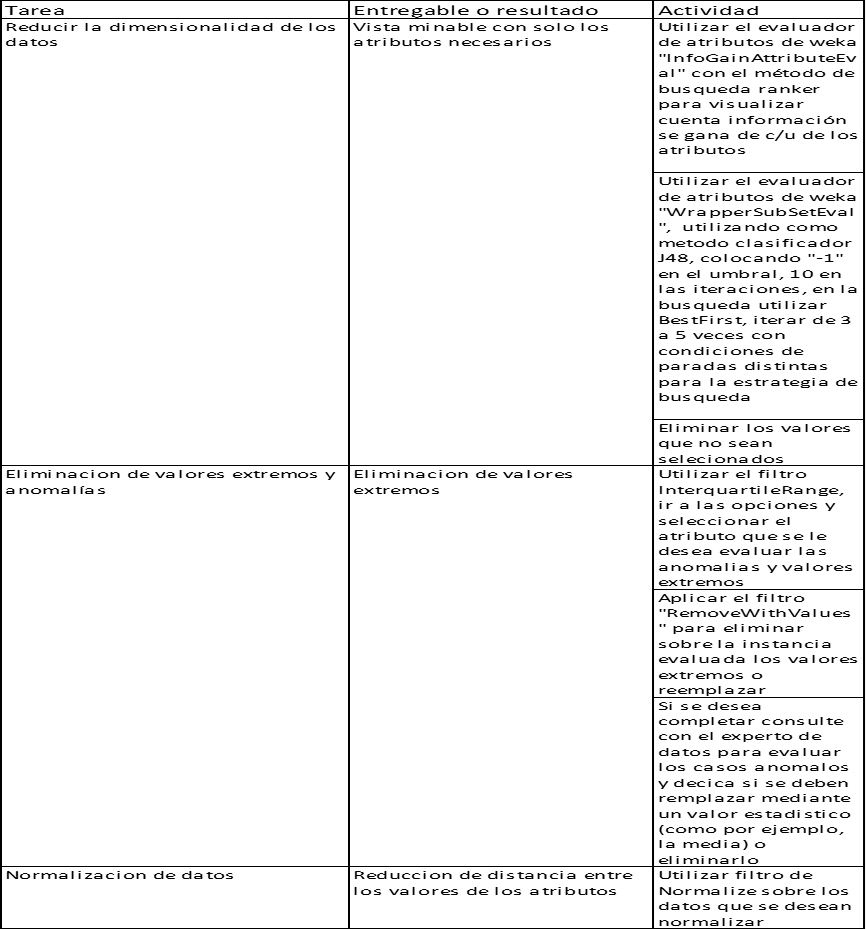


**Anexo 4 Reglas basándose en conocimiento adquirido de exploración de los datos**





**Anexo 5 Tareas entregables y actividades de pre procesamiento por K medias**





**Anexo 6 Reglas basándose en conocimiento adquirido para pre procesamiento de datos con algoritmo K medias**





**Anexo 7 Tareas, entregables y actividades para pre procesamiento sobre algoritmo a priori**



**Anexo 8 Reglas basándose en conocimiento adquirido de pre procesamiento para algoritmo a priori**

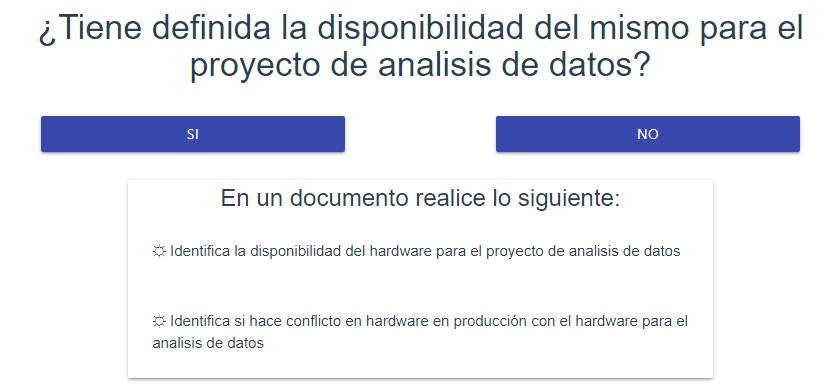




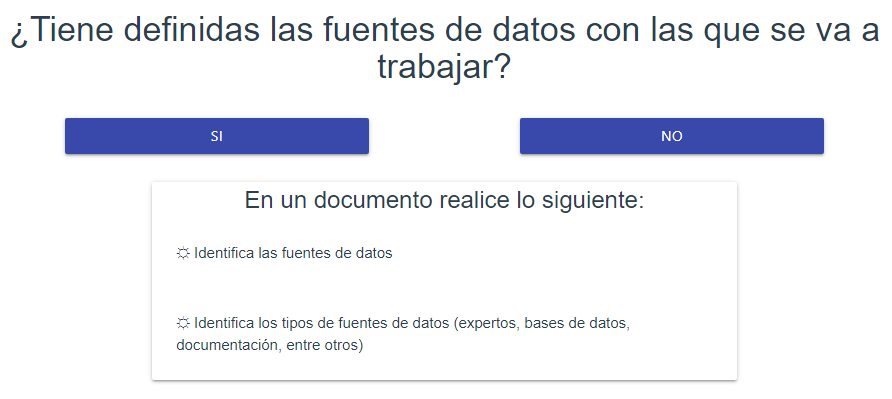
Anexo 9 Reglas de aplicación de algoritmo a priori y k medias



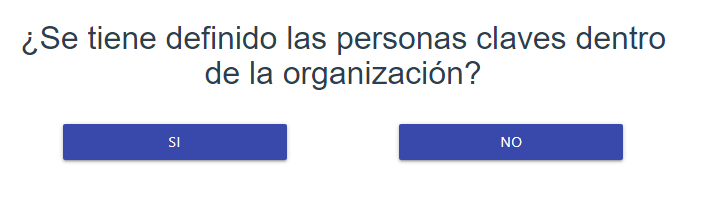
**Anexo 10 Pregunta acerca de la disponibilidad de los datos realizada por el asistente durante etapa 1 del proceso de análisis de datos**



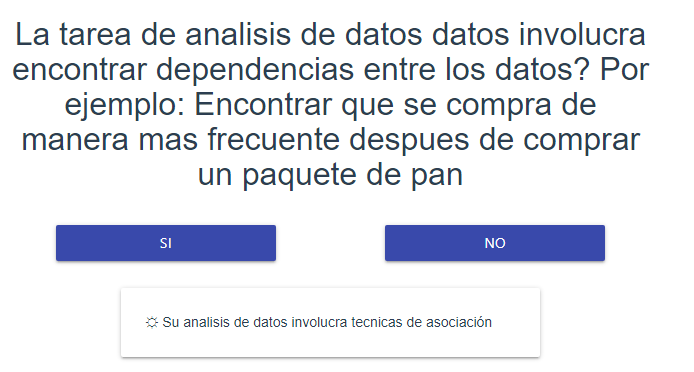
**Anexo 11 Pregunta acerca de la definición de las fuentes de datos realizada por el asistente durante la etapa 1 del proceso de análisis de datos**

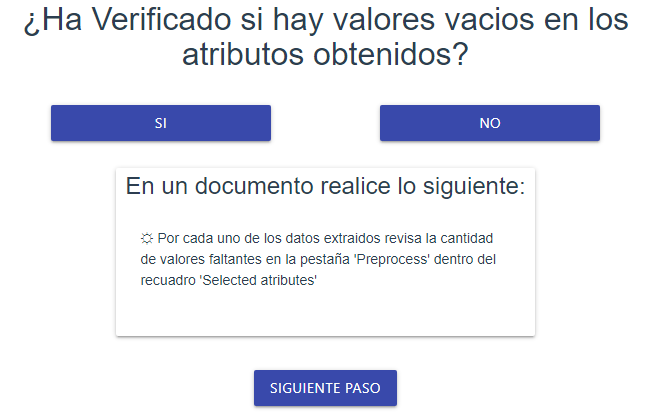


**Anexo 12 Pregunta acerca de la organización realizada por el asistente durante la etapa 1 del proceso de análisis de datos**



**Anexo 13 Recomendación del asistente acerca del tipo de técnica que debe ser utilizada en el proceso de análisis de datos**





**Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos**

**Anexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos**

**Anexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos**

**Anexo 15Anexo 14 Recomendación del asistente al dar una respuesta negativa acerca del conocimiento del usuario en sus datos acerca de valores vacíos en sus datos**

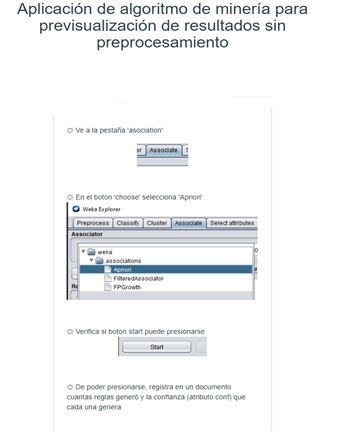
**Anexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori**

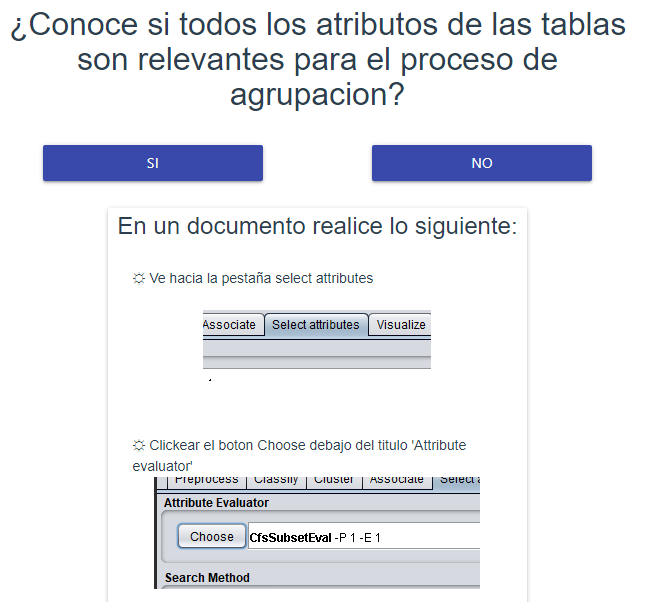
**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de característicasAnexo 15 Fragmento de pantalla de recomendación del asistente acerca de cómo aplicar algoritmo a priori**





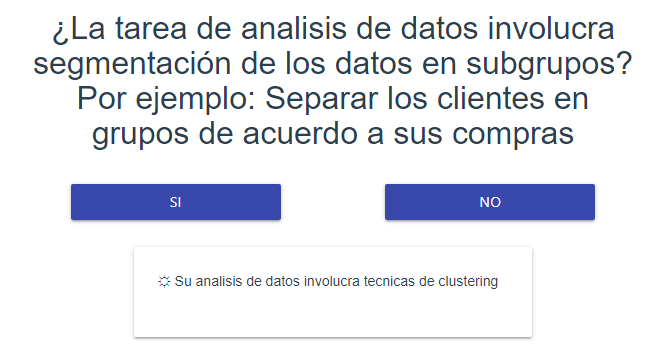
**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características**

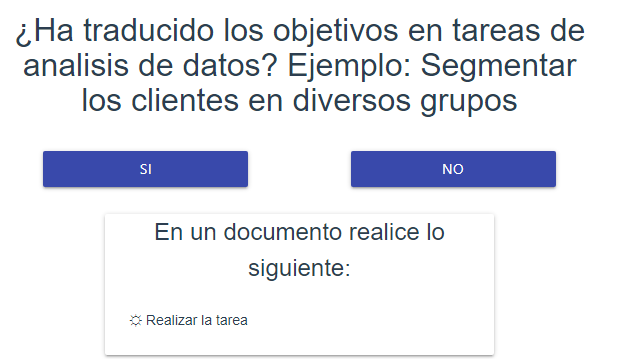
**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características**

**Anexo 16 Fragmento de pantalla del asistente sobre cómo aplicar algoritmo de reducción de características**

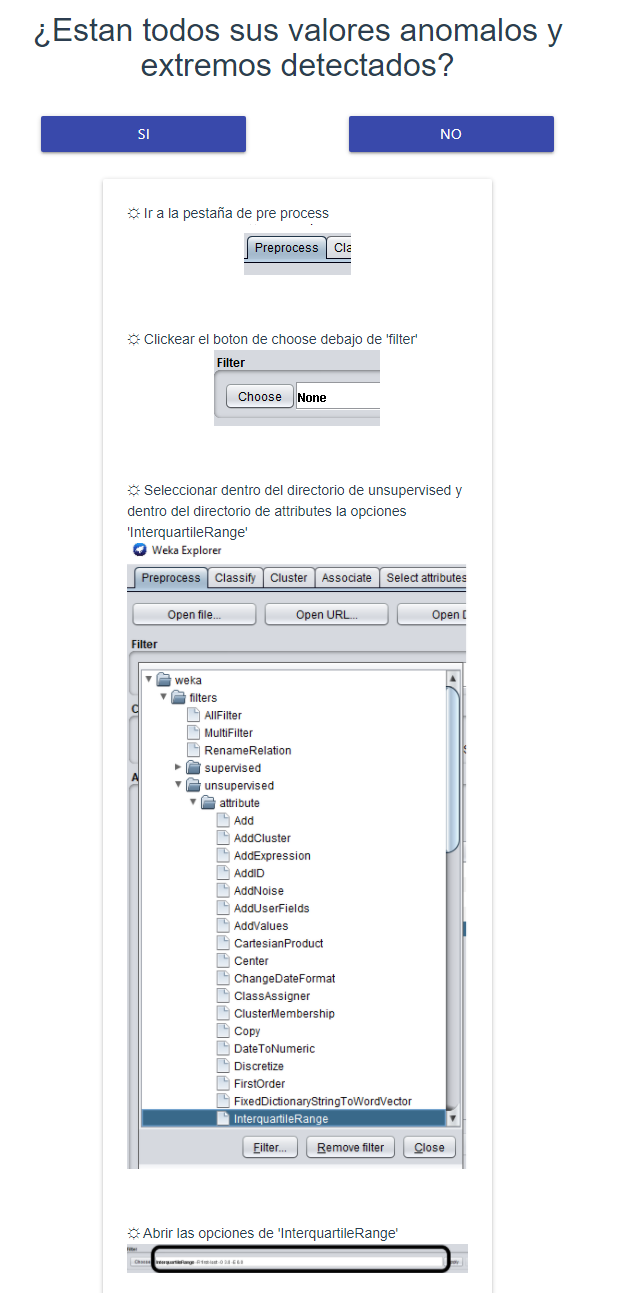
**Anexo 17 Recomendación del asistente para utilizar técnicas de agrupamiento o clustering**



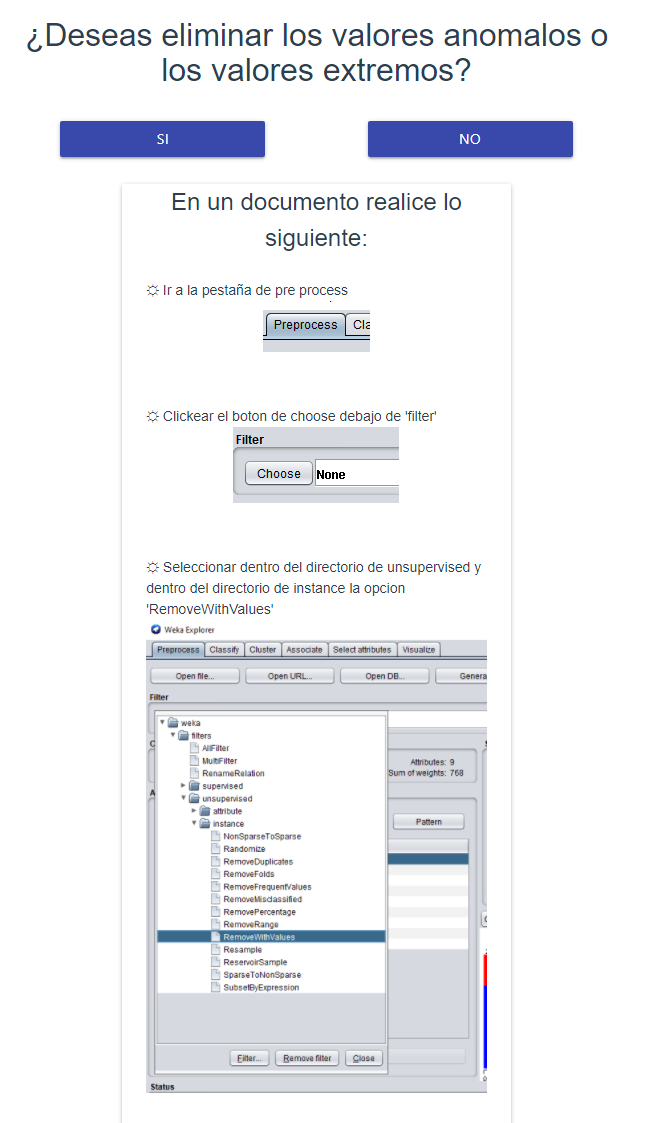
**Anexo 18 Recomendación del asistente acerca de puntos a tener en cuenta para en análisis**



**Anexo 19 Fragmento de recomendación del asistente acerca de cómo detectar valores anómalos y extremos**



**Anexo 20 Fragmento de pantalla del asistente acerca de como eliminar valores anómalos**



**Anexo 21 Fragmento de pantalla del asistente acerca de como aplicar el algoritmo de k medias**

