

Tesis de Maestría

MINERÍA DE DATOS COMO FACTOR ESTRATÉGICO PARA GENERAR

VENTAJAS COMPETITIVAS APLICADO AL ÁREA DE RR.HH. DE LAS EMPRESAS EN LA REPUBLICA ARGENTINA

Ing. Kenny Fernand Huertas Noreña

FACULTAD DE INGENIERÍA

Maestría en Tecnología de la Información

Director de Tesis

Prof. Dr. Nicolás D’Ippolito

Buenos Aires, Argentina 2018



Facultad de Ingeniería

MINERÍA DE DATOS COMO FACTOR ESTRATÉGICO PARA GENERAR

VENTAJAS COMPETITIVAS APLICADO AL ÁREA DE RR.HH. DE LAS EMPRESAS EN LA REPUBLICA ARGENTINA

Ing. Kenny Fernand Huertas Noreña

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Director de Tesis

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Presidente del Jurado

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jurado

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jurado Externo

Defensa de la Tesis

Ciudad Autónoma de Buenos Aires, a los \_\_\_ días del mes de\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

De \_\_\_\_\_\_\_\_

II

*A mi familia, por su apoyo incondicional*

*en este largo viaje académico. A mis compañeros y amigos.*

III

RECONOCIMIENTOS

A mi director de tesis, Nicolás D’Ippolito, por apoyarme en la investigación y por ser mi guía en el proceso académico.

A mis profesores de la Maestría en Tecnología de la Información de la Universidad de Palermo por dar nuevas perspectivas en la carrera de la ingeniería de sistemas.

A mis compañeros de la maestría y por compartir conocimientos al son de una cerveza.

IV

TABLA DE CONTENIDO

[1.1 INTRODUCCIÓN 1](#_Toc108643)

[1.2 DESCRIPCIÓN DEL PROBLEMA 3](#_Toc108644)

[1.2.1 CUÁL ES EL PROBLEMA 3](#_Toc108645)

[1.2.2 SUPOSICIONES 4](#_Toc108646)

[1.2.3 PROBLEMAS SIMILARES A LA INVESTIGACION 4](#_Toc108647)

[1.2.4 PORQUÉ EL PROBLEMA DEBE SER RESUELTO 5](#_Toc108648)

[1.2.5 COMO RESOLVERÍA EL PROBLEMA 5](#_Toc108649)

[1.3 OBJETIVO GENERAL 5](#_Toc108650)

[1.3.1 OBJETIVOS ESPECÍFICOS 6](#_Toc108651)

[1.4 ALCANCES Y LIMITACIONES 6](#_Toc108652)

[1.5 JUSTIFICACIÓN 7](#_Toc108653)

[2. MARCO TEÓRICO 9](#_Toc108654)

[2.1 CRITERIOS DE ÉXITO 9](#_Toc108655)

[2.1.1 ANÁLISIS FODA 11](#_Toc108656)

[2.2 GESTORES DE BASE DE DATOS (SGBD) 12](#_Toc108657)

[2.2.1 SISTEMAS SQL 13](#_Toc108658)

[2.2.2 SISTEMAS ORIENTADOS A OBJETOS 14](#_Toc108659)

[2.2.3 SISTEMAS NOSQL (2000) 15](#_Toc108660)

[2.2.4 SISTEMAS XML 16](#_Toc108661)

[2.2.5 LENGUAJES DE MODELACIÓN 16](#_Toc108662)

[2.3 ESTANDAR 20](#_Toc108663)

[2.3.1 ANSI/X3/SPARC 20](#_Toc108664)

[2.3.2 MODELOS DE REFERENCIA DE ANSI 23](#_Toc108665)

[2.3.3 ODMG 25](#_Toc108666)

[2.3.4 CONCLUSIÓN 26](#_Toc108667)

[2.4 CLASIFICACION DE BASES DE DATOS 27](#_Toc108668)

[2.4.1 BODEGA DE DATOS (DATA WAREHOUSE) 31](#_Toc108669)

[2.4.2 DEFINICIÓN DE DATAWAREHOUSE 35](#_Toc108670)

[2.4.3 PARADIGMA BILL INMON 36](#_Toc108671)

[2.4.4 PARADIGMA RALPH KIMBALL 38](#_Toc108672)

[2.4.5 DIMENSIONAL MODELING (DM) 41](#_Toc108673)

[2.4.6 DEFINICIÓN DE ETL 45](#_Toc108674)

[2.5 MINERÍA DE DATOS 52](#_Toc108675)

[2.5.1 OBJETIVO DE MINERÍA DE DATOS 52](#_Toc108676)

[2.5.2 TIPOS DE MINERÍA DE DATOS 58](#_Toc108677)

[2.5.3 FUNCIONES Y TAREAS DE LA MINERÍA DE DATOS 60](#_Toc108678)

[2.5.3 TÉCNICAS DE MINERÍA DE DATOS 77](#_Toc108679)

[2.5.4 TIPOS DE MODELOS PARA ANÁLISIS PREDICTIVOS 85](#_Toc108680)

[2.5.5 HERRAMIENTAS DE ANÁLISIS PREDICTIVO 89](#_Toc108681)

[3. METODOLOGÍA DE LA INVESTIGACIÓN 99](#_Toc108682)

[3.1 METODOLOGÍAS DE MINERÍA DE DATOS 99](#_Toc108683)

[3.1.1 METODOLOGÍA SEMMA 100](#_Toc108684)

[3.1.2 METODOLOGÍA KDD 102](#_Toc108685)

[3.1.3 METODOLOGÍA CRISP-DM 105](#_Toc108686)

[3.2 COMPARACIÓN DE METODOLOGÍAS 110](#_Toc108687)

[3.3 RECURSOS HUMANOS 112](#_Toc108688)

[4. DISEÑO DE LA SOLUCIÓN 113](#_Toc108689)

[4.1 ANÁLISIS DE DATOS 115](#_Toc108690)

[4.2 PREPARACIÓN DE LOS DATOS 116](#_Toc108691)

[4.2.1 PROCESO KDD KNOWLEDGE DISCOVERY IN DATABASES 117](#_Toc108692)

[4.3 CREACIÓN DE LA BASE DE DATOS 119](#_Toc108693)

[4.4 CONSTRUCCION DEL ETL 120](#_Toc108694)

[4.4.1 PREPARAR LOS DATOS ETL 121](#_Toc108695)

[4.4.2 SELECCIÓN DE LA FUENTE DE DATOS 121](#_Toc108696)

[4.4.3 CONSTRUCCION DEL DATAWAREHOUSE 128](#_Toc108697)

[4.5 DISEÑO DE PRUEBAS 130](#_Toc108698)

[4.5.1 DETECCION DE PATRONES 131](#_Toc108699)

[4.5.2 EMPLEADO ACTUAL 135](#_Toc108700)

[4.5.3 NUEVO EMPLEADO 141](#_Toc108701)

[4.5.3 GRÁFICAS CONSOLIDADO 151](#_Toc108702)

[4.6 ANALISIS DE COSTOS 157](#_Toc108703)

[5. RESULTADOS Y CONCLUSIONES 160](#_Toc108704)

[5.1 RESULTADOS 162](#_Toc108705)

[5.2 CONCLUSIONES 163](#_Toc108706)

[5.3 TRABAJO FUTURO 166](#_Toc108707)

[6. BIBLIOGRAFÍA 166](#_Toc108708)

V

VI

VII

LISTADO DE IMÁGENES

Imagen 1. Matriz Análisis FODA .................................................................................................. 11

Imagen 2. Gestores de base de datos (SGBD) ............................................................................... 14

Imagen 3. Estructura jerárquica usada en los SGBD ..................................................................... 18

Imagen 4. Modelo en red. .............................................................................................................. 19

Imagen 5. Estructura relacional ..................................................................................................... 20

Imagen 6. Estructura multidimensional ......................................................................................... 21

Imagen 7. Forma dimensional. ....................................................................................................... 34

Imagen 8. Paradigma Bill Inmon ................................................................................................... 40

Imagen 9. Arquitectura DW ........................................................................................................... 41

Imagen 10. Modelo Estrella ........................................................................................................... 45

Imagen 11. Modelo Snowflake o copo de nieve ............................................................................ 45

Imagen 12. Modelo Constelación .................................................................................................. 46 Imagen 13. Proceso ETL ................................................................................................................ 49

Imagen 14. Asociación diversas disciplinas de minería de datos. ................................................. 57

Imagen 15. Tipología de los algoritmos de minería de datos ........................................................ 65 Imagen 16. Pseudocódigo del algoritmo KNN .............................................................................. 68

Imagen 17. Búsqueda de clases a la que pertenece la clasificación del punto. .............................. 68

Imagen 18. Concepto de margen e hiperplano ............................................................................... 70

Imagen 19. Ecuación de Centroide ................................................................................................ 76

Imagen 20. Asignación de los centroides....................................................................................... 77

Imagen 21. Ecuación de Centroide ................................................................................................ 77

Imagen 22. Ecuación de Centroide ................................................................................................ 78

Imagen 23. Microsoft-machine-learning-algorithm-cheat-sheet-v6 .............................................. 81

Imagen 24. Proceso de extracción de patrones mediante el uso de técnicas de minería de datos. . 82

Imagen 25. Árbol de decisión, ejemplo entrevista laboral ............................................................. 88

Imagen 26. Logo Python ................................................................................................................ 96

Imagen 27. Opciones de gráficos Matplot ..................................................................................... 98

Imagen 28. Libro Python for Data Analysis .................................................................................. 99

Imagen 29. Framework Bottle ..................................................................................................... 100

Imagen 30. Logo de Jupyter......................................................................................................... 101

Imagen 31. Distribución Anaconda ............................................................................................. 102

Imagen 32. Base de datos PostgreSQL ........................................................................................ 103

Imagen 33. Talend Studio ............................................................................................................ 103 Imagen 34. Etapas de KDD ......................................................................................................... 108

Imagen 35. Proceso de descubrimiento de conocimiento en bases de datos (KDD). .................. 110

Imagen 36. Metodología CRISP-DM .......................................................................................... 111

Imagen 37. Esquema de los cuatro niveles de abstracción de la metodología CRISPDM. ......... 112

Imagen 38. Encuesta realizada por la KDnuggets en el año 2007 ............................................... 116

Imagen 39. Proceso KDD ........................................................................................................... 124

Imagen 40. Fuente de datos ......................................................................................................... 128 Imagen 41. Diseño físico de la tabla ............................................................................................ 129

Imagen 42. Carpeta datos a unificar............................................................................................. 130

Imagen 43. Diagrama Copo de nieve ........................................................................................... 135

Imagen 44. Construcción del DataWarehouse ............................................................................. 136

Imagen 45. Elección de un motivo de renuncia ........................................................................... 142

Imagen 46. Elección de un objetivo Laboral ............................................................................... 143

Imagen 47. Empleado por días identificando el motivo de renuncia ........................................... 144

Imagen 48. Datos del empleado con el motivo de renuncia y con la sugerencia a tomar ............ 145

Imagen 49. Elección de tipo renuncia y objetivo laboral ............................................................. 146

Imagen 50. Gráfico empleado por días identificando el motivo de renuncia .............................. 147

Imagen 51. Librerías propias de Scikit-learn ............................................................................... 148

Imagen 52. Validación *train\_test\_split* ........................................................................................ 148

Imagen 53. Matriz de asignación de campos ............................................................................... 149

Imagen 54. Asignación de variable *Predict* ................................................................................. 149

Imagen 55. Porcentaje Obtenido .................................................................................................. 150

Imagen 56. Porcentaje Obtenido .................................................................................................. 150

Imagen 57. Porcentaje Obtenido .................................................................................................. 150

Imagen 58. Datos del candidato ingresados en el formulario ...................................................... 152

Imagen 59. Resultado que predijo el algoritmo ........................................................................... 153

Imagen 60. Caso de estudio para otro candidato ......................................................................... 154

Imagen 61. Resultado que predijo el algoritmo ........................................................................... 155

Imagen 62. Caso de estudio para otro candidato ......................................................................... 156

Imagen 63. Resultado que predijo el algoritmo ........................................................................... 157

Imagen 64. Elección de tipo renuncia y objetivo laboral ............................................................. 158

LISTADO DE TABLAS

Tabla 1. Esquemas de la Arq. de sistemas de bases de datos......................................................... 22

Tabla 2. Niveles de la arquitectura................................................................................................. 23

Tabla 3. Objetivos del Modelo de Referencia ................................................................................ 25

Tabla 4. Técnicas orientadas a objetos........................................................................................... 26 Tabla 5. Estándar ODMG .............................................................................................................. 27

Tabla 6. Relación existente entre las tareas y técnicas de minería de datos. ................................. 89

Tabla 7. Analogía entre las metodologías KDD y CRISP-DM ................................................... 117

Tabla 8. Factores de deserción con valores de la empresa. .......................................................... 121

Tabla 9. Campos disponibles para la preparación de datos ......................................................... 131

Tabla 10. Empleados por periodo ................................................................................................ 138

Tabla 11. Ganancia por empleado................................................................................................ 164 Tabla 12. Ganancia por empleado................................................................................................ 165

LISTADO DE ILUSTRACIONES

Ilustración 1. Tablas sexo, tipo\_objetivo, tipos\_de\_renuncia ...................................................... 126

Ilustración 2. Carga del proyecto ................................................................................................. 132

Ilustración 3. Configuración con las fuentes de datos.................................................................. 133

Ilustración 4. Conexión con las fuentes de datos ......................................................................... 133

Ilustración 5. Prueba de conexión ―Exitoso‖ con las fuentes de datos ........................................ 134

Ilustración 6. Conexión con las fuentes de datos ......................................................................... 134

Ilustración 7. Salida del proceso .................................................................................................. 136

LISTADO DE GRÁFICOS

Gráfico 1. Motivos de renuncia: Abandono de trabajo y Renuncia ............................................. 159

Gráfico 2. Motivos de renuncia: Abandono de trabajo, Renuncia y Fin de contrato a prueba ... 160

Gráfico 3. Motivos de renuncia: Fin de contrato a prueba ........................................................... 161

Gráfico 4. Motivos de renuncia: Renuncia y Fin de contrato a prueba ....................................... 162

XII

Í

## 1.1 INTRODUCCIÓN

Los saltos en el almacenamiento y el poder computacional de la última década nos llevan a invertir y aprovechar la información de los trabajadores, utilizando bodegas de datos y aplicando las diferentes técnicas de extracción de conocimiento usando minería de datos (modela la base de datos con el propósito de determinar patrones y relaciones entre los datos) dado que son capaces de proporcionar una mejor comprensión de los diversos datos existentes, procesos de negocio o la vida cotidiana de una manera que no era posible en el pasado.

Con los resultados de estas técnicas podemos llegar a examinar los datos relevantes que se pueden presentar, con el fin de entender mejor el trabajo de los empleados de la empresa y los contextos en que ellos se desenvuelven.

Lo cual, quiere decir que los datos a considerar hoy en día en cualquier tipo de empresa ya no son solo de tipo texto, con la revolución de la imagen, sonido, el crecimiento de internet y el continuo desarrollo de tecnologías hace que la información cada vez sea más compleja tanto en su estructura interna como semántica (entiéndase por semántica a los aspectos del significado, sentido o interpretación de signos lingüísticos como símbolos, palabras, expresiones). Los tiempos están cambiando y también lo hacen los requisitos del negocio y las exigencias de los usuarios se han vuelto aún más complejas.

La gran cantidad de datos obtenidos día a día en las empresas proporcionan amplias oportunidades pero estos mismos vienen con desafíos significativos en términos de gastos computacionales y de inversión en el almacenamiento de la información. Las técnicas de extracción de información han sido empleadas con éxito para crear modelos, patrones, tendencias y algoritmos de predicción en diferentes contextos, obteniendo resultados prometedores que demuestran cómo determinadas características sociológicas, económicas, educativas, etc. pueden afectar el rendimiento laboral encontrando un indicador que permita identificar a los empleados con mayor riesgo de fallo o abandono, juicios laborales, rendimiento laboral, entre otros.

Es importante predecir la posibilidad que esto suceda desde que ingresa a la empresa y poder cambiar los factores que pudieran estar causando ésta situación, puesto que generalmente no se les da el seguimiento adecuado y no se hace un diagnóstico a tiempo antes que el trabajador lo haga. Utilizando la minería de datos se puede calcular el porcentaje de probabilidad y de esta manera se propondrán con anticipación las estrategias necesarias para disminuir el porcentaje de los factores que inciden en el empleado.

Schermerhorn J.M en su artículo ―Management. Hoboken: Wiley (2012)‖ Citado en el capítulo 6 *Methodological Aspects of Information Sharing and Management in*

*Organizations* en el libro *Secure Information Management Using Linguistic Threshold Approach* de la editorial *Springer* propone que ―La información que se encuentra en una empresa de cualquier tipo constituye su enorme activo, fortaleza y ventaja, puede darle una posición que lo distinga de otras empresas similares en la jerarquía y determinar adecuadamente su lugar en el mercado (…) mientras que la falta de una definición de información significativa dentro de una organización representa su debilidad‖.

## 1.2 DESCRIPCIÓN DEL PROBLEMA

Para la identificación del problema se utiliza de apoyo la metodología CRISP-DM. En esta fase se comprende que es lo que realmente necesita la empresa a través de la construcción del prototipo. Esta fase se incluye la identificación del problema a resolver, restricciones y suposiciones, problemas similares, como también los objetivos de la investigación, y, finalmente, producir el plan del proyecto.

### 1.2.1 CUÁL ES EL PROBLEMA

Se necesita un programa que informe qué tipo de renuncia tendrán los empleados de la empresa. Revisar miles de empleados y seleccionar los mejores para optimizar la búsqueda de posibles candidatos a renunciar.

Tom Mitchell[[1]](#footnote-1) propone una decisión precisa para un programa informático, dice que aprende de la experiencia *E* con respecto a alguna clase de tareas *T* y medida de rendimiento *P*, si su desempeño en tareas en *T*, medido por *P*, mejora con la experiencia *E*. Esto sería:

* Tarea (T): clasificar un empleado que no ha sido catalogado dentro del set de prueba.
* Experiencia (E): un/unos empleado(s) de la empresa que ya tienen un tipo de renuncia establecido y de otros no.
* Rendimiento (P): Exactitud de clasificación, el número de renuncia predichas correctamente.

### 1.2.2 SUPOSICIONES

* Las variables específicas utilizadas en el set de pruebas son importantes para el modelo.
* El usuario de recursos humanos no le importa cómo funciona el

algoritmo/modelo.

* La cantidad de registros puede ser importante para el modelo.
* Los registros antiguos y nuevos son de igual importancia para predecir correctamente.

### 1.2.3 PROBLEMAS SIMILARES A LA INVESTIGACION

* Abandono de clases escolares, alumnos que se retiran escuela/universidad.
* Clientes que se van de compañías de cable y telefonía.
* Detección y clasificación para la correcta selección de candidatos a un empleo.
* Detección de correos de publicidad y ventas que pueden ser catalogados y clasificados como spam mediante filtros de palabras de un algoritmo de selección.

### 1.2.4 PORQUÉ EL PROBLEMA DEBE SER RESUELTO

* Motivación: aprender de un algoritmo de aprendizaje junto con técnicas y herramientas de Data Mining.
* Surge como nivel académico para obtener un grado de Maestría en Tecnología De La Información.
* Usar este proyecto con beneficios laborales a empresas que sean del rubro de recursos humanos.
* Expandir el proyecto para más compañías y analizar diferentes rubros laborales.

### 1.2.5 COMO RESOLVERÍA EL PROBLEMA

 Extrayendo la información de diferentes fuentes de datos, procesarla en una base de datos, transformarla mediante procesos ETL y usar minería de datos para obtener patrones usando algoritmos de Machine Learning para posterior mostrarla los en un entorno web.

## 1.3 OBJETIVO GENERAL

Diseñar y generar mediante la utilización de aplicaciones de Data Mining un modelo de minería de datos para aislar e identificar patrones, tendencias y las variables de los datos sensibles y realmente importantes que afectan a los trabajadores de la empresa.

### 1.3.1 OBJETIVOS ESPECÍFICOS

* Diseñar y crear una bodega de datos, aplicando técnicas de ETL a partir de múltiples fuentes de información.
* Proveer al área de recursos humanos de las organizaciones las variables que afectan el desempeño de los trabajadores.
* Seleccionar, aplicar y evaluar un proceso para la búsqueda de patrones de comportamiento usando los datos de la empresa con métodos de Machine

Learning, de esa forma obtener y generar mediante un algoritmo y modelo No Supervisado un prototipo funcional de minería de datos que sea útil para la toma de decisiones.

* Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones de representación y de inteligibilidad.

## 1.4 ALCANCES Y LIMITACIONES

La información para implementar el modelo de minería de datos será proporcionada por el área de recursos humanos y el área administrativa de la empresa *Argentina de*

*Seguridad*. En la empresa existe un registro de más de 11 años y cuenta con más de 5000 registros. Al iniciar cada año, depende de la cantidad de nuevos proyectos contratados pueden ingresan en promedio 180 empleados de lo contrario pueden egresan un promedio de 230 empleados con relación al año pasado. Esta información será

relevante para poder lograr los objetivos de la tesis.

## 1.5 JUSTIFICACIÓN

El interrogante acerca de las personas y su relación con el mundo de trabajo, las condiciones de vida, calidad del entorno laboral, mala relación entre el empleado y el jefe directo, injusticia en la retribución salarial, los estilos de adaptación de cada trabajador son algunos de los aspectos y factores que generan un impacto directo en la estabilidad y desempeño laboral del trabajador manifestándose en un alto índice de rotación de personal para la empresa. El presente trabajo busca mediante el análisis de la aplicación de diferentes técnicas y herramientas de minería de datos la identificación de patrones y secuencias de comportamiento, algoritmos y modelos de búsqueda de datos, con el fin de predecir anomalías, características y múltiples factores que puedan influir en el día a día de los trabajadores.

La gestión del área de recursos humanos es la de buscar e incorporar para un puesto de trabajo a una persona que cumpla los requisitos técnicos y académicos sin importar su lugar de vivienda, a veces no influye su nivel académico. Pero el lugar de vivienda es uno de los principales factores y variables determinantes ante el panorama de la deserción laboral, cuestiones de salud y juicios laborales juegan también un papel importante y los empleados que tienen que viajar grandes distancias para llegar al trabajo son buenos candidatos para el retraso excesivo, las ausencias y, finalmente, salir a buscar otra opción laboral más cercana a su residencia.

Hoy en día las empresas en especial las pequeñas pymes reconocen que los datos que poseen son un recurso valioso para crecer en el mercado y competir contra las demás. Los datos deben ser administrados y gestionados por estándares que cumplan reglas de privacidad y de negocio. El uso y la implementación de soluciones y aplicaciones de técnicas de minería de datos *(Data Mining)* da lugar al descubrimiento de información y a la adquisición de conocimiento útil a partir de sus bases de datos permitiendo el acceso a una innumerable cantidad de información que está revolucionando el área de RRHH y es una tarea fundamental para la empresa, el transformar los datos en información y la información en inteligencia para el desarrollo, reconocimiento y evaluación de los empleados. El área de Recursos Humanos es una de las más importantes y principales de toda organización donde se gestiona en todo momento diferentes tipos de esfuerzos que van dirigidos a los colaboradores de la empresa para un mejoramiento eficiente que ayude a alcanzar directa o indirectamente las metas de los clientes.

Hace 10 años atrás, Alejandro Melamed director general de *Humanize Consulting* definió a esta área con un rol crítico para establecer el propósito y el sentido de la organización. Nuestra capacidad de capturar, almacenar y comprender cantidades masivas de datos está cambiando la ciencia, la medicina, los negocios y la tecnología así lo declara la sección especial sobre "*La Edad de Petabyte*" en la edición de la revista Wired del mes de Junio del año 2008. Hoy esta definición no ha cambiado y cada vez más las empresas le apuestan a ésta area.

Por lo tanto, el foco de este trabajo a través de herramientas y técnicas de Data Mining se basará en identificar variables, características y patrones de comportamiento relacionados a los ámbitos laborales.

# 2. MARCO TEÓRICO

## 2.1 CRITERIOS DE ÉXITO

Al cumplir con los objetivos se podrá determinar cuáles empleados necesitarán más apoyo o algún seguimiento por parte del área de recursos humanos para evitar su deserción laboral. Para este correcto análisis se deben definir las fuerzas competitivas del sector de la empresa de caso de estudio. Se detallan dichas fuerzas:

* Rivalidad del sector empresarial o Competidores en el sector o Equilibrio entre competidores o Crecimiento del sector empresarial o Costos del cambio de proveedor
* Nuevos competidores en el mercado o Curva de aprendizaje o Diferenciación de producto o Necesidades de capital
* Amenaza de productos parecidos o Existencia de sectores similares o Propensión del cliente al cambio o Relación calidad/precio
* Poder de negociación de la empresa

o Por costos se produzca cambios en el proveedor del servicio de vigilancia o Integración/compra de grandes empresas a pequeñas pymes

* Poder de negociación de los clientes o Rentabilidad y margen de los clientes o Grado de conocimiento de la empresa en el sector

### 2.1.1 ANÁLISIS FODA

Para la investigación se hace uso de la matriz foda, es una herramienta de estudio para evidenciar la situación actual de una empresa, institución, proyecto o persona, se hace un análisis de las características internas *(Debilidades y Fortalezas)* y externas *(Amenazas y*

*Oportunidades)* en una matriz cuadrada.

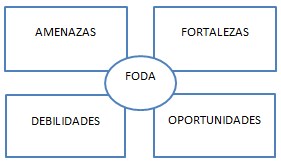
Imagen

1

. Matriz Análisis FODA

Fuente: E

laboración propia



* Fortalezas o Buen soporte comercial o Amplio número de recursos o Más de 120 clientes con la empresa o Mayor cobertura en el sector empresarial
* Debilidades o Servicio costoso o Calidad baja o Retrasos en los turnos o Ausencias importantes en los reemplazos o Alto índice de rotación
* Oportunidades o Buena política gubernamental de ayudas a las empresas o Economía en auge o Mercado abierto o Buenas posibilidades de crecimiento
* Amenazas o Aumento de competidores o Inestabilidad de los sueldos en el sector empresarial o Directivas del gobierno

## 2.2 GESTORES DE BASE DE DATOS (SGBD)

Los gestores de base de datos (SGBD) son programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos dentro del nivel conceptual, dentro de los programas de software que manejan las bases de datos proporcionan herramientas para añadir, borrar, modificar y analizar los datos. Codd, E. F.

(1970) en el artículo publicado por el ACM *"A Relational Model of Data for Large Shared Data Banks"* dice que los usuarios pueden acceder a la información mediante consultas SQL y dependiendo el repositorio usar filtros o preguntas específicas de interrogación para la generación de informes. Las bases de datos han estado en uso desde los primeros días de las computadoras electrónicas.

A diferencia de los sistemas modernos, que se pueden aplicar a datos y necesidades muy diferentes, la mayor parte de los sistemas originales estaban enfocados a bases de datos específicas y pensadas para ganar velocidad a costa de perder flexibilidad. Según Juan

Iruela (2016, enero 19) define a los gestor de base de datos *(DataBase Managenent System)* como un sistema que permite la creación, gestión y administración de bases de datos, así como la elección y manejo de las estructuras necesarios para el almacenamiento y búsqueda de la información del modo más eficiente posible. En la actualidad, existen multitud de SGBD en la mayoría relacionales.

Imagen 2. Gestores de base de datos (SGBD)



Fuente: Elaboración propia

### 2.2.1 SISTEMAS SQL

IBM comenzó a trabajar a principios de 1970 en un prototipo lejanamente basado en los conceptos de *Codd* llamándolo System R. La primera versión estuvo lista en 1974 y comenzó así el trabajo en sistemas multi-tabla, en los que los datos podían disgregarse de modo que toda la información de un registro (alguna de la cual es opcional) no tuviera que estar almacenada en un único segmento grande. Las versiones multiusuario siguientes fueron aprobadas por los usuarios en 1978 y 1979, tiempo en que el lenguaje

SQL había sido estandarizado. En Suecia, el artículo de Codd generó la base de datos Mimer SQL en la universidad de Uppsala. En 1984 este proyecto se consolidó en una compañía independiente. A principios de 1980, Mimer. (s. f.). Introdujo la gestión de transacciones para dar robustez a las aplicaciones, una idea que fue recogida en muchos otros SGBD.

### 2.2.2 SISTEMAS ORIENTADOS A OBJETOS

Durante la década de 1980 el auge de la programación orientada a objetos influyó en el modo de manejar la información de las bases de datos. Programadores y diseñadores comenzaron a tratar los datos en las bases de datos como objetos. Esto quiere decir que si los datos de una persona están en la base de datos, los atributos de la persona como dirección, teléfono y edad se consideran que pertenecen a la persona, no son datos extraños. Esto permite establecer relaciones entre objetos y atributos, más que entre campos individuales. Otro gran foco de atención durante la década fue el incremento de velocidad y fiabilidad en el acceso.

La idea consistía en replicar la información importante y más solicitada en una base de datos temporal de pequeño tamaño con enlaces a la base de datos principal. Esto implicaba que se podía buscar mucho más rápido en la base de datos pequeña que en la grande. Su mejora de prestaciones llevó a la introducción de la indización, incorporada en la totalidad de los SGBD.

### 2.2.3 SISTEMAS NOSQL (2000)

El siglo XXI trajo una nueva tendencia en las bases de datos, el NoSQL. Esta tendencia introducía una línea no relacional significativamente diferente de las clásicas. No requieren por lo general esquemas fijos, evitan las operaciones *JOIN* almacenando datos des-normalizados y están diseñadas para escalar horizontalmente.

La mayor parte de ellas pueden clasificarse como almacenes clave-valor o bases de datos orientadas a documentos. Como dice Nancy Lynch (2002) Recientemente ha habido una gran demanda de bases de datos distribuidas con tolerancia a particiones, pero el teorema CAP *(consistencia, disponibilidad y tolerancia a las particiones)* que consiste básicamente en que todos los nodos vean la misma información simultáneamente y que a su vez garanticen la correcta confirmación de la petición de la invocación a los mismos.

El autor Eric Brewer en (Brewer, 2000) presenta la *Conjetura De Brewer* también llamada de esta forma y dice que no es posible conseguir un sistema distribuido que simultáneamente proporcione consistencia, disponibilidad y tolerancia al particionado. Un sistema distribuido puede satisfacer sólo dos de las tres restricciones a la vez. Por dicha razón muchas de las bases de datos NoSQL usan la llamada consistencia eventual para proporcionar disponibilidad y tolerancia al particionado, con un nivel máximo de consistencia de datos.

Entre las aplicaciones más populares de código abierto encontramos:

* MongoDB
* MemcacheDB
* RedisCouchDB
* Hazelcast
* Apache Cassandra
* HBase

### 2.2.4 SISTEMAS XML

Las Bases de Datos XML forman un subconjunto de las Bases de Datos NoSQL. Todas ellas usan el formato de almacenamiento XML, que está abierto, legible por humanos y máquinas y ampliamente usado para interoperabilidad. En esta categoría encontramos:

* BaseX
* eXist
* MarkLogic Server
* MonetDB/XQuery
* Sedna

### 2.2.5 LENGUAJES DE MODELACIÓN

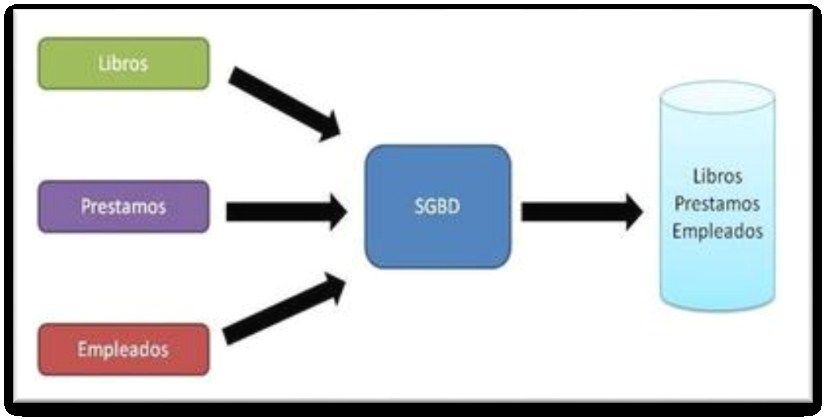
Toda base de datos soportada por un SGBD debe tener esquemas modelados

adecuadamente. Coincidiendo con la evolución histórica de las bases de datos, estas han utilizado distintos modelos. Los SGBD esperan un modelo determinado para poder acceder de forma simple a la base de datos. Estos modelos son:

* Jerárquicos
* En red
* Relacionales
* Multidimensionales

La estructura jerárquica fue usada en los SGBD de los primeros mainframe.

Imagen 3. Estructura jerárquica usada en los SGBD



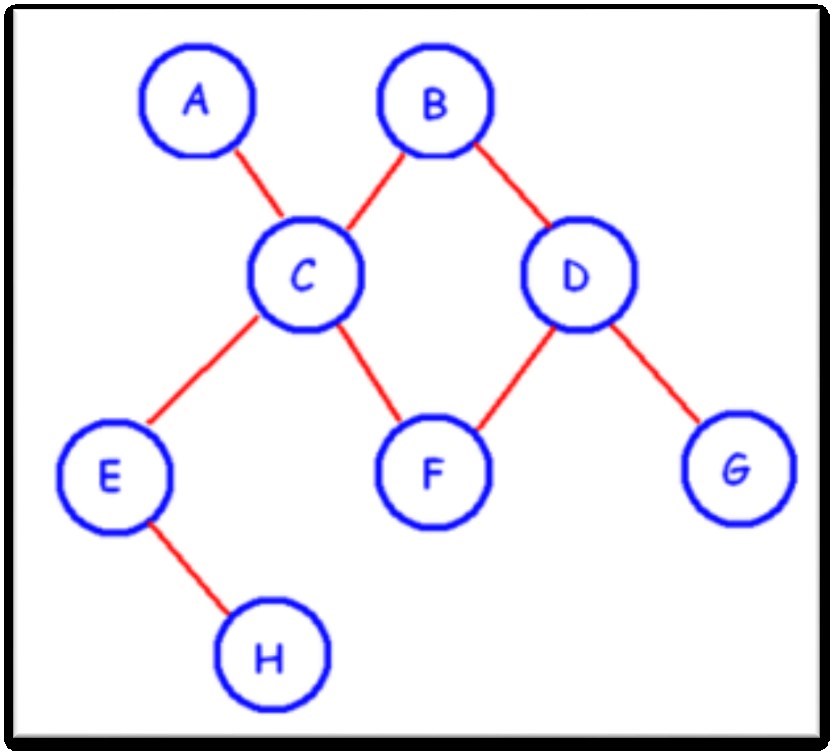
Fuente: Elaboración propia

Esta estructura es simple pero inflexible ya que las relaciones están confinadas al tipo 1:n. El sistema IMS de IBM y el RDM Mobile de Raimason (2018) da ejemplos de bases de datos con múltiples jerarquías sobre el mismo conjunto de datos.

En cambio la estructura en red obtiene relaciones más complejas que las jerárquicas.

Admite relaciones de cada registro con varios que se pueden seguir por distintos caminos. En otras palabras, el modelo permite relaciones N:N. El modelo en red está concebido como un modo flexible de representar objetos y sus relaciones. Su cualidad distintiva es que el esquema visto como un conjunto de nodos conectados por arcos no tiene ninguna restricción.

Imagen 4. Modelo en red.



Fuente: Elaboración propia

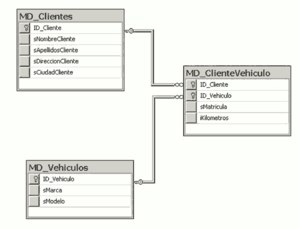
El inventor de este modelo fue Charles Bachman, y el estándar fue publicado en 1969 por

CODASYL.

La estructura relacional que definió E. F. Codd (1972) es la más extendida hoy en día en IBM en 1972, esta estructura se usa en mainframes, computadoras medias y

microcomputadoras. Almacena los datos en filas (tuplas) y columnas (atributos).

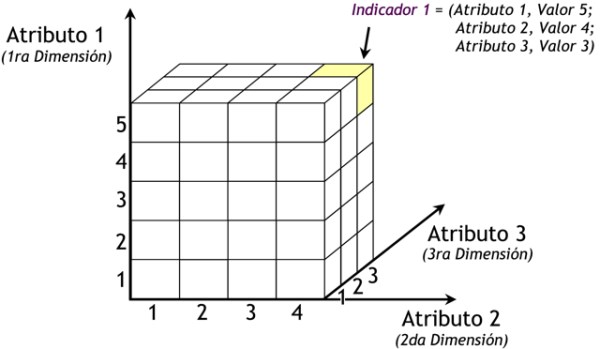
Imagen 5. Estructura relacional



Fuente: Elaboración propia

La estructura multidimensional tiene parecidos a la del modelo relacional, pero en vez de las dos dimensiones filas-columnas, tiene N dimensiones. Esta estructura ofrece el aspecto de una hoja de cálculo. Es fácil de mantener y entender ya que los registros se almacenan del mismo modo como se detalla en la *imagen 4*. Sus altas prestaciones han hecho de ella la base de datos más popular para el proceso analítico de transacciones en línea (OLAP).

Imagen 6. Estructura multidimensional



Fuente: Adaptado de 3.4 Datawarehouse manager | Manual Dataprix TI. (s. f.).

Recuperado el 24 de enero de 2018, a partir de [http://www.dataprix.com/datawarehouse-](http://www.dataprix.com/datawarehouse-manager)

[manager](http://www.dataprix.com/datawarehouse-manager)

## 2.3 ESTANDAR

### 2.3.1 ANSI/X3/SPARC

ANSI/X3/SPARC (American National Standards Institute)[[2]](#footnote-2), está dentro del Comité X3, que se ocupa de ordenadores e informática. Básicamente estudian los

Sistemas de Gestión de Bases de Datos en el ANSI y tiene relación con ISO (International Standards Organization). En el informe que se generó en 1975 se discutió la generación de una arquitectura de un sistema de gestión de bases de datos que incluían la identificación y descripción de sus múltiples interfaces. La finalidad de este informe era presentar un marco para el análisis de la arquitectura y de sus interfaces.

Como ayuda para poder conseguir la separación entre los programas de aplicación y los datos y como se almacenaba el esquema de la base de datos la ANSI-SPARC estableció tres esquemas dentro de la arquitectura de sistemas de bases de datos.

Tabla 1. Esquemas de la Arq. de sistemas de bases de datos.

|  |  |
| --- | --- |
| ESQUEMAS | OBJETIVO/DESCRIPCION |
| NIVEL  INTERNO | Tiene un esquema interno que describe la estructura física de almacenamiento de base de datos. Emplea un modelo físico de datos y los únicos datos que existen están realmente en este nivel. |
| NIVEL  CONCEPTUAL | Tiene esquema conceptual. Describe la estructura de toda la base de datos para una comunidad de usuarios. Oculta los detalles físicos de almacenamiento y trabaja con elementos lógicos como entidades, atributos y relaciones |
| NIVEL  EXTERNO | Tiene varios esquemas externos o vistas de usuario. Cada esquema describe la visión que tiene de la base de datos a un grupo de usuarios, ocultando el resto. |

Fuente: Elaboración propia

En este modelo la arquitectura de tres niveles juega un papel importante ya que establece el concepto de independencia de datos; definiendo como se modifica el esquema en un nivel del sistema sin tener que modificar el esquema del nivel superior.

Tabla 2. Niveles de la arquitectura

|  |  |
| --- | --- |
| TIPOS DE  INDEPENDENCIA | OBJETIVO/DESCRIPCION |
| LÓGICA | Capacidad de modificar el esquema conceptual sin tener que alterar los esquemas externos ni los programas de aplicación. Como modificar el esquema conceptual para ampliar la base de datos o para reducirla. |
| FÍSICA | Capacidad de modificar el esquema interno sin tener que alterar el esquema conceptual (o los externos). Mejorar el rendimiento de las operaciones de consulta o de actualización de datos. Dado que la independencia física se refiere sólo a la separación entre las aplicaciones y las estructuras físicas de almacenamiento, es más fácil de conseguir que la independencia lógica. |

Fuente: Elaboración propia

En conclusión el uso de esta arquitectura ANSI/SPARC es uno de los conceptos clave ya que responde a las exigencias de evolución del sistema de información, ayudando a una mejor utilización de los recursos y las inversiones realizadas.

### 2.3.2 MODELOS DE REFERENCIA DE ANSI

Se definió en 1982, el DAFTG (*Datábase Architecture Framework Task Group*) [[3]](#footnote-3)marco de trabajo para arquitecturas de bases de datos. Este grupo describe objetivos que el Modelo de Referencia pretende alcanzar, al presentar un marco común para la descripción de los SGBD, facilito su estudio y análisis de forma sistemática ayudando en la labor de estandarización generando la compatibilidad de los distintos componentes de los SGBD.

El informe[[4]](#footnote-4) presentó un modelo de referencia (MR) para la estandarización de los Sistemas Gestores de Bases de Datos, fue publicado en marzo, 1986. Según dicho informe hace referencia a una estructura conceptual que facilitaba el trabajo de estandarización de las BD, identificando una serie de componentes y viendo cómo se interrelacionan. Los objetivos que debe cumplir el MR son: la adaptación al desarrollo tecnológico, nuevas arquitecturas, unificación de los modelos de datos, compatibilidad con otros modelos de referencia y estándares.

Al cumplir estos objetivos se gana beneficios en la portabilidad de las aplicaciones y sobre todo en la productividad de las empresas que es lo que realmente buscan los altos gerentes y áreas interesadas.

Tabla 3. Objetivos del Modelo de Referencia

|  |  |
| --- | --- |
| PUNTOS DE VISTA | OBJETIVO |
| Componentes-  integración | Busca la compatibilidad entre los módulos de un SGBD y a su vez de compatibilidad en el mercado. |
| Funcionesespecificación | Buscar que parámetros se utilizan para pasar datos e información de control y cómo van a tener la ejecución de las funciones dentro de los SGBD. |
| Datos- utilizar y  describir | * Como se pueden entonces manipular (insertar, borrar, modificar y seleccionar) los datos por medio de un lenguaje de manipulación, como SQL. * Establecer una petición de datos por parte del usuario siendo ejecutada por los transformadores conceptual/externo, interno/conceptual y almacenamiento/interno, que utilizan los metadatos. |
| Arquitectura  ANSI/SPARC | Describir las interrelaciones del SGBD básicamente se ocupa del qué, por qué y para qué de las mismas. |
| Interfaz de lenguaje de datos (LD) | Permitir a los usuarios y a los procesadores especificar sus peticiones para la recuperación de los datos por el SGBD. |
| Interfaz lenguaje de datos internos | Permitir el uso de los servicios de los procesadores que soportan el funcionamiento de los SGBD, en especial los del SO. |

Fuente: Elaboración propia

### 2.3.3 ODMG

El objetivo de este modelo fue el de permitir que los diseños como las implementaciones fueran portables entre los sistemas que lo soportan, al mismo tiempo especificó los elementos que se tenían que definir y de qué manera se debía hacer. Consta de un lenguaje de definición de objetos, ODL, que especifica los elementos de este modelo. El estándar ODMG principalmente proporciona técnicas *orientadas a objetos* para la ingeniería de software.

Básicamente define la semántica de los objetos de una base de datos, a su vez permite la portabilidad, interoperabilidad de los diseños como de las implementaciones en diversos sistemas permitiendo que una aplicación se pueda ejecutar con mínimas modificaciones.

Tabla 4. Técnicas orientadas a objetos

|  |  |  |
| --- | --- | --- |
| OBJETO |    | Instancia de una entidad de interés.  Identificador único (OID- Identificador de Objeto). |
|  |  | Valor específico. |
|  |  | No tiene que ser necesariamente un solo valor. |
| LITERAL |  | Puede ser una estructura o un conjunto de valores relacionados  que se guardan bajo un solo nombre (por ejemplo,  enumeraciones). |

Fuente: Elaboración propia

El estándar ODMG está compuesto por:

Tabla 5. Estándar ODMG

|  |  |  |
| --- | --- | --- |
| OML: Lenguaje de  manipulación de objeto |  | Declaración de nuevas clases: Database, Transaction, Session, Conection. |
|  |  | Nuevos objetos: bajo el Operador "New(bd) nombre\_clase" |
| ODL: Lenguaje de  definición de objeto |  | Define los atributos y las relaciones entre tipos. ODL es el equivalente de DDL (lenguaje de definición de datos) de los  SGBD relacionales. |
| OQL: Consulta de  objetos |  | Proporciona un acceso declarativo a los objetos. Tiene una sintaxis muy similar a SQL |

Fuente: Elaboración propia

### 2.3.4 CONCLUSIÓN

El uso de un estándar en los SGDB es uno de los conceptos clave ya que responde a las exigencias de evolución del sistema de información, ayudando a una mejor utilización de los recursos (humanos y de máquina) y preservando las inversiones realizadas; de este modo adaptando al sistema tecnológico nuevas arquitecturas al unificar los modelos de datos, modelos de referencia y estándares generando beneficios en la portabilidad de las aplicaciones y en la productividad de las empresas.

## 2.4 CLASIFICACION DE BASES DE DATOS

Las bases de datos que usan y que implementan en la gestión del manejo de datos hoy en día en las grandes empresas y medianas también por sus características y funcionalidades que estas proponen se pueden clasificar bajo un estándar o marco de referencia en el sector de sistemas. Cabe destacar que el almacenamiento de información y grandes volúmenes en la manipulación de datos juega un papel importante en la toma de decisiones. Se definirá los conceptos claves de los gestores y motores de datos en este documento como también la clasificación dependiendo el uso de las mismas.

Una base de datos es una colección de archivos interrelacionados, creados con un sistema manejador de bases de datos. El contenido que engloba a la información concerniente (almacenadas en archivos) de una organización, de tal manera que los datos estén disponibles para los usuarios.

Una finalidad es eliminar la redundancia o al menos minimizarla. Petrovic, D. (2010) comenta de los tres componentes principales de un sistema en base de datos son el hardware, el software DBMS y los datos a manejar, así como el personal encargado del manejo del sistema.

Datos por acá, datos por allá, documentos con información valiosa en cajas, contratos encima del escritorio y la gente necesita tener análisis de datos de forma rápida y muchas veces no sabe dónde buscar o guardar la información. Los datos son los más importantes al momento de la toma de decisiones, los datos que se obtienen en el día a día puede ser clasificados dentro de dos grupos: los *no estructurado* y *estructurados*. Estos mismos con gran contenido valioso para el ciclo de vida de la empresa ya que las personas interesadas consiguen respuestas efectivas por el análisis de métricas, informes y ventas con tendencias, todo esto gracias a manipular la información que se guarda en las bases datos o en repositorios.

Nacen nuevas oportunidades de negocio con la utilización y la implementación de tecnologías para manipular los datos no estructurados, definimos este grupo como documentos con imágenes y texto, archivos PDF, hojas de cálculo, imágenes, videos, archivos de audio, presentaciones con diapositivas, publicaciones en medios sociales. Estos últimos mediante técnicas como BigData y DataWarehouse facilitan que las organizaciones en las altas gerencias puedan analizar con informes reales la situación exacta de la empresa, tarea, proyecto y no justamente gracias a la utilización de consultas a bases de datos relacionadas o con comandos SQL.

Las bases de datos pueden clasificarse[[5]](#footnote-5) de acuerdo al uso que se va a dar a los datos, a la información que se quiera manipular.

* Bases de datos Estáticas:

Son base de datos de sólo lectura se utilizan mucho para realizar ejecución en la toma de decisiones y realizar análisis de datos, por ejemplo base de datos de empresas del rubro periódicos ya que almacenan información por si se requiere consultarla tiempo después.

* Bases de datos Dinámicas:

Estas son base de datos donde la información almacenada se modifica con el tiempo, permiten que se haga una actualización, borrado y adición de datos, además de las operaciones fundamentales de consulta. Son utilizadas en muchos ámbitos de negocios, en un sistema de información de un supermercado, una farmacia.

* Bases de datos Centralizadas:

Es muy similar al modelo de cliente/servidor en el sentido que la base de datos está centralizada en un lugar y los usuarios están distribuidos.

Este modelo sólo brinda la ventaja de tener el procesamiento distribuido ya que en sentido de disponibilidad y fiabilidad de los datos no se gana nada.

* Base de datos Distribuidos:

Una base de datos distribuida (BDD) es un conjunto de múltiples bases de datos que se encuentran distribuidas en diferentes espacios lógicos (por ejemplo, un servidor corriendo 2 máquinas virtuales) e interconectados por una red de comunicaciones. Dichas BDD tienen la capacidad de realizar procesamiento autónomo, esto permite realizar operaciones locales o distribuidas. Un sistema de Bases de Datos Distribuida (SBDD) es un sistema en el cual múltiples sitios de bases de datos están ligados por un sistema de comunicaciones de tal forma que, un usuario en cualquier sitio puede acceder los datos en cualquier parte de la red exactamente como si estos fueran accedidos de forma local. Esto permite que un usuario en cualquier sitio pueda acceder a los datos en cualquier parte de la red exactamente como si éstos fueran accedidos de forma local. Un ejemplo sería el uso dentro de una empresa una base de datos Oracle 11g o Microsoft SQL Server 2008 las computadoras reciben el nombre de nodos.

* Base de datos Federada:

Una base de datos Federada es un sistema en el que varias bases de datos parecen funcionar como una sola entidad. Cada componente de la base de datos en el sistema es completamente auto sostenido y funcional. Cuando una aplicación consulta la base de datos el sistema analiza cuál de los componentes de la base contiene los datos que se solicitan y pasa la solicitud a la misma. SQL Server es un ejemplo de una base de datos federada.

* Base de datos NoSQL:

También llamadas *no solo SQL, son* un enfoque hacia la gestión de datos y el diseño de base de datos para grandes volúmenes de información, abarca una amplia gama de tecnologías y arquitecturas, gracias a esto busca resolver los problemas de escalabilidad y rendimiento de BigData que las bases de datos relacionales no fueron diseñadas para abordar.

Las grandes y pequeñas empresas gracias a esta característica acceden y analizan grandes cantidades de datos no estructurados o datos que se almacenan de forma remota en varios servidores virtuales en la nube, en escritorios, cajones de oficina, cajas o dispositivos de almacenamiento tipo USB. Las bases de datos más populares NoSQL son Apache Cassandra, simpleDB, Google BigTable, Mapreduce, Memcachedb y Voldemort. Las empresas que utilizan NoSQL incluyen NetFlix, LinkedIn y Twitter. Por ejemplo, pares Clave/valor o tuplas apoyándose del conocimiento de los metadatos, también se usa formato UML para el almacenamiento de la información contenida en los datos no estructurados.

### 2.4.1 BODEGA DE DATOS (DATA WAREHOUSE)

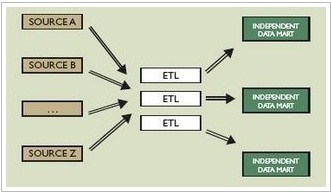
Una bodega de datos lo define Torres-Jiménez, J. (2011) como un conjunto de datos integrados u orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales apoyan el proceso de toma de decisiones de la administración y está orientada al manejo de grandes volúmenes de datos provenientes de diversas fuentes o diversos

tipos.

El concepto de DataWarehouse (DW) llegó de la mano de Bill Inmon y Ralph Kimball (2013). Ambos pensaron en un único repositorio de información para poder integrar y explotar información de diversos sistemas fuentes. Pero, más allá de esta generalización conceptual, cada uno decidió hacerlo a su manera.

Entonces, veamos qué es lo que propone cada uno de ellos: Inmon propone un enfoque Top-Down en donde Los datos son extraídos de los sistemas operacionales por los procesos ETL y consolidados en el Datawarehouse corporativo, mientras que Kimball (2013) sugiere utilizar una metodología Bottom-Up, donde la información se extrae de los sistemas transaccionales para ser cargada en diferentes Data Marts cada uno de los cuales son independientes, están modelados en forma dimensional y tienen foco departamental.

Imagen 7. Forma dimensional.



Fuente: Adaptado de Alexander Factor (2001). Analyzing Application Service Providers. Prentice Hall PTR. p.290

Estos Data Marts podrían ser implementados con arquitecturas ROLAP o MOLAP.

2.4.1.1 SISTEMAS ROLAP

La arquitectura ROLAP, accede a los datos almacenados en un Datawarehouse para proporcionar los análisis OLAP. El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. Esta arquitectura accede directamente a los datos del Data Warehouse haciendo que las consultas que ejecutan los usuarios finales sean más rápidas ya que soportan múltiples *Joins* y relaciones entre múltiples tablas.

Entre sus ventajas, esta arquitectura se considera más escalable para manejar grandes volúmenes de datos. Otra ventaja es que hay disponible en el mercado una gran variedad de herramientas de carga de datos para este sistema. Los datos en estas bases de datos pueden ser accedidas por cualquier herramienta de generación de informes y consultas SQL (reporting). Como desventaja las herramientas ROLAP no disponen de mecanismos automáticos para la carga de las tablas desde sistemas ETL, lo que significa más tiempo en el desarrollo de código. Las técnicas de compresión de información no estan buena como en sistemas Molap haciendo que el almacenamiento en disco crezca

2.4.1.2 SISTEMAS MOLAP

Por sus siglas en ingles *procesamiento analítico multidimensional en línea.* Es una arquitectura de dos niveles que usa bases de datos multidimensionales y motores analíticos para proporcionar el respectivo análisis de la información de la empresa. La base de datos multidimensional es la encargada del manejo, acceso y obtención de los datos. Para calcular estos mismos y la información agregada procedente de los sistemas operacionales, se carga mediante una serie de rutinas batch a través de las dimensiones del negocio. Entre sus ventajas, gracias a la optimización del rendimiento de almacenamiento, la indexación multidimensional y la memoria caché las consultas son más rápidas a la tablas de datos. A diferencia de los sistemas Rolap, los procesos de ETL están mejor detallados y automatizados mejorando los tiempos de desarrollo.

Como desventaja se tiene esto se consigue que los tiempos de carga sean generalmente mucho menores que con las cargas MOLAP automatizadas. Esta arquitectura no es muy considera paras ser utilizada de forma escalable puesto que para manejar grandes volúmenes de datos requiere mayores tiempos en procesar y cargar la información y en conseguir una herramienta en el mercado que lo soporte.

2.4.1.3 COMPARATIVA ROLAP-MOLAP

A continuación se hace una comparativa de los dos sistemas:

* El sistema Rolap delega la negociación entre tiempo de respuesta y el proceso batch al diseño del sistema. Mientras, que el sistema Molap, requiere que sus bases de datos estén pre compiladas para conseguir un rendimiento mejor en las consultas, incrementando, por tanto los requerimientos batch.

* Los sistemas con alto índice al cambio constante de los datos requieren una arquitectura que pueda realizar estas modificaciones en tiempo real. Los sistemas ROLAP soportan de forma estable y dinámica esta arquitectura, mientras que los sistemas MOLAP están más orientados hacia consolidaciones batch en un tiempo determinado.

### 2.4.2 DEFINICIÓN DE DATAWAREHOUSE

DataWarehouse o almacén de datos, es definido como un proceso de gestión centralizada de datos, en donde se aplican herramientas que permiten resumir, describir y analizar los datos históricos para dar soporte a la toma de decisiones (Sumathi y Sivanandam, 2006). Un DataWarehouse proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios. Normalmente en el almacén de datos habrá que guardar información histórica que cubra un amplio período de tiempo. Pero hay ocasiones en las que no se necesita la historia de los datos, sino sólo sus últimos valores, siendo además admisible generalmente un pequeño desfase o retraso sobre los datos operacionales.

Podemos entender un DataMart como un subconjunto de los datos del DataWarehouse con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Al igual que en un DataWarehouse, los datos están estructurados en modelos de estrella o copo de nieve y un DataMart puede ser dependiente o independiente de un DataWarehouse.

El DataMart está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización. Tal como lo define Jill Dyché (2000). El DataWarehouse es la organización en su conjunto es el almacén natural para los datos corporativos comunes.

### 2.4.3 PARADIGMA BILL INMON

Bill Inmon[[6]](#footnote-6) ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis (sería el CIF o Corporate Information Factory). Insiste además en que ha de tener las siguientes características:

* Orientado a temas:

Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.

* Integrado:

La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.

* No volátil:

La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.

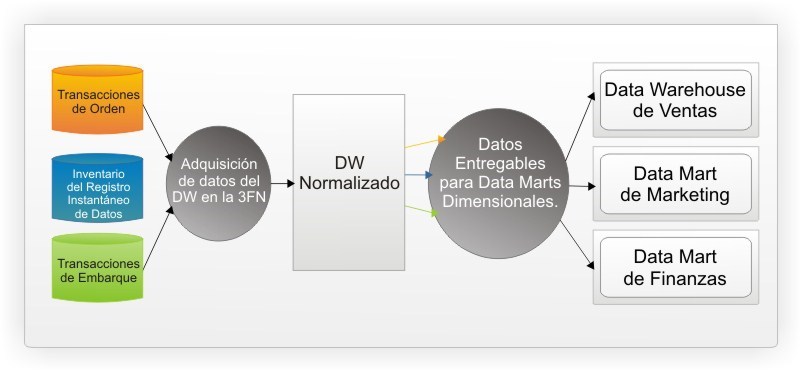
* Variante en el tiempo:

Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

La información ha de estar a los máximos niveles de detalle. Los DW departamentales o DataMarts son tratados como subconjuntos de este DW corporativo, que son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir de este DW Central (del que también se pueden construir los ODS (Operational *Data Sores*) o similares que estén dentro de la capa de integración como servidores de homologación o áreas de staging). Se entiende *ODS* como contenedores de datos en uso. Este mismo puede integrar muchos datos de varias fuentes de información en uno solo, siendo este su propósito principal.

Para el tratamiento de los cambios en los datos, usa la *Continúe and Discrete Dimension Management* (inserta fechas en los datos para determinar su validez para las Continue Dimension o bien mediante el concepto de snapshot o foto para las Discrete Dimension). Al tener este enfoque global, es más difícil de desarrollar en un proyecto sencillo (pues estamos intentando abordar el ―todo‖, a partir del cual luego iremos al ―detalle‖).

Imagen 8. Paradigma Bill Inmon



Fuente: Adaptado de Alexander Factor (2001). Analyzing Application Service Providers.

Prentice Hall PTR. p.290.

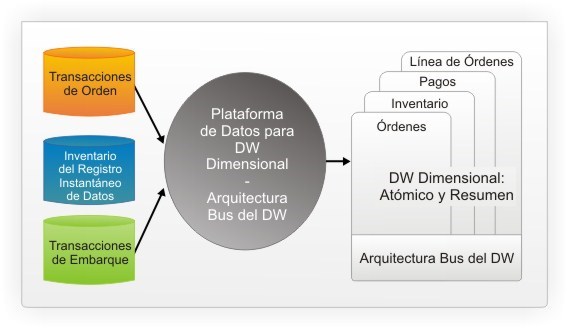
### 2.4.4 PARADIGMA RALPH KIMBALL

El DataWarehouse es un conglomerado de todos los DataMarts según manifiesta Kimball (2013) dentro de una empresa, siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional (no normalizado), que incluye, como ya vimos, las dimensiones de análisis y sus atributos, su organización jerárquica, así como los diferentes hechos de negocio que se quieren analizar. Por un lado tenemos tablas para las representar las dimensiones y por otro lado tablas para los hechos (las *facts tables*).

Los diferentes DataMarts están conectados entre sí por la llamada *bus structure*, que contiene los elementos anteriormente citados a través de las dimensiones conformadas

(que permiten que los usuarios puedan realizar querys conjuntos sobre los diferentes DataMarts, pues este bus contiene los elementos en común que los comunican). Una dimensión conformada puede ser, por ejemplo, la dimensión cliente, que incluye todos los atributos o elementos de análisis referentes a los clientes y que puede ser compartida por diferentes DataMarts (ventas, pedidos, gestión de cobros, etc.).

Imagen 9. Arquitectura DW



Fuente: Adaptado de Alexander Factor (2001). Analyzing Application Service Providers.

Prentice Hall PTR. p.290.

Este enfoque también se referencia como Bottom-up, pues al final el DataWarehouse Corporativo no es más que la unión de los diferentes DataMarts, que están estructurados de una forma común a través de la bus structure.

Esta característica le hace más flexible y sencillo de implementar, pues podemos construir un DataMart como primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL[[7]](#footnote-7) extraen la información de los sistemas operacionales y los procesan igualmente en el *area stage (área de almacenamiento)*, realizando posteriormente el llenado de cada uno de los DataMart de una forma individual, aunque siempre respetando la estandarización de las dimensiones (dimensiones conformadas).

La metodología para la construcción del DW incluye 4 fases:

* Selección del proceso de negocio.
* Definición de la granularidad de la información
* Elección de las dimensiones de análisis
* Identificación de los hechos o métricas.

Se define el tratamiento de los cambios en los datos a través de las Dimensiones Lentamente Cambiantes (SCD). Kimball, Ralph; Margy Ross (2002) llaman a la metodología "Dimensional Modeling (DM)", o también conocida como "La Metodología

Kimball", ha llegado a convertirse en un estándar en el área del soporte de decisiones. Una de las técnicas favoritas en la construcción de un DataWarehouse, en la cual se constituyen modelos de tablas y relaciones con el propósito de optimizar la toma de decisiones con base en mediciones de los resultados de los procesos de negocio.

### 2.4.5 DIMENSIONAL MODELING (DM)

Técnica de diseño lógico que tiene como objetivo presentar los datos dentro de un marco de trabajo estándar e intuitivo, para permitir su acceso con un alto rendimiento. Cada Modelo Dimensional[[8]](#footnote-8) está compuesto por una tabla con una llave combinada, llamada tabla de hechos, y un conjunto de tablas más pequeñas llamadas tablas de dimensiones.

Los elementos de estas tablas se pueden definir de la siguiente manera:

* Hechos:

Colección de piezas de datos y datos de contexto. Cada hecho representa una parte del negocio, una transacción o un evento.

* Dimensiones:

Una dimensión es una colección de miembros, unidades o individuos del mismo tipo.

* Medidas:

Son atributos numéricos de un hecho que representan el comportamiento del negocio relativo a una dimensión.

Entre las principales fortalezas que según comenta Whitehorn, Mark (2010), para este modelo se pueden mencionar las siguientes:

* Resistencia a cambios en la conducta del usuario.

El diseño lógico de este modelo puede realizarse independientemente de los patrones esperados de consulta, pues todas las dimensiones son generadas como puntos de entrada simétricos a la tabla de hechos.

* Flexible para aceptar datos nuevos e inesperados.

La tabla de hechos y las tablas de dimensiones pueden ser alteradas simplemente agregando los registros de los nuevos datos en la tabla correspondiente. La información no tiene que ser cargada nuevamente. Las herramientas para generar reportes o consultas no necesitan ser reprogramadas para adaptarse a los cambios. Y, finalmente, las aplicaciones utilizadas continúan ejecutándose sin cambios en su rendimiento.

Se denomina modelo estrella o esquema de estrella-unión a un modelo dimensional como una tabla de hechos rodeada por las tablas de dimensiones.

Imagen 10. Modelo Estrella



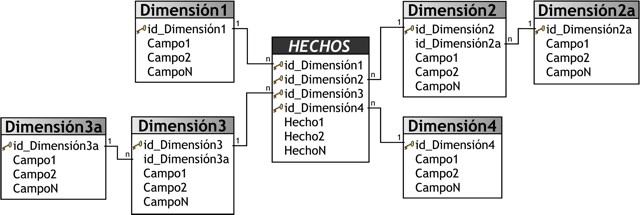
Fuente: Adaptado de D3.4 Datawarehouse manager | Manual Dataprix TI. (s. f.).

Recuperado el 24 de enero de 2018, a partir de [http://www.dataprix.com/datawarehouse-](http://www.dataprix.com/datawarehouse-manager)

[manager](http://www.dataprix.com/datawarehouse-manager)

Otra variante es la que se conoce como Snowflake o copo de nieve, en donde se presentan ramificaciones a partir de las tablas de dimensiones y no sólo a partir de la tabla de hechos.

Imagen 11. Modelo Snowflake o copo de nieve



Fuente: Adaptado de 3.4 Datawarehouse manager | Manual Dataprix TI. (s. f.).

Recuperado el 24 de enero de 2018, a partir de [http://www.dataprix.com/datawarehouse-](http://www.dataprix.com/datawarehouse-manager)

[manager](http://www.dataprix.com/datawarehouse-manager)

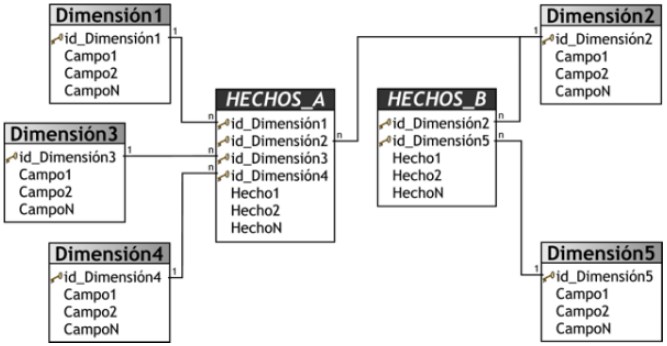
El esquema constelación se puede construir con base en los dos modelos anteriores, este contiene múltiple tablas de hechos y es muy complejo de utilizar ya que las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos. A continuación se detalla el modelo descrito.

Imagen

12

.

Modelo Constelación



Fuente: Adaptado de 3.4 Datawarehouse manager | Manual Dataprix TI. (s. f.).

Recuperado el 13 de julio de 2018, a partir de [http://www.dataprix.com/datawarehouse-](http://www.dataprix.com/datawarehouse-manager)

[manager](http://www.dataprix.com/datawarehouse-manager)

Pese a las diferencias que se pueden apreciar de manera inmediata entre ambas arquitecturas, también existen elementos en común. Todas las empresas requieren almacenar recursos, analizar e interpretar la información que generan y acumulan con el fin de tomar decisiones críticas que permitan su existencia pero sobre todo que maximicen su prosperidad; por lo que se vuelve prioritario crear sistemas de análisis y retroalimentación para comprender su información (DataWarehouse) y de esta manera contar con los elementos adecuados para la toma de decisiones. Otra característica del DataWarehouse es que contiene metadatos, es decir, datos sobre los datos. Los metadatos permiten saber la procedencia de la información, su periodicidad de actualización, su fiabilidad, forma de cálculo, etc.

Cada punto de entrada a la tabla de hechos está conectado a una dimensión, lo que permite determinar el contexto de los hechos. Una base de datos dimensional se puede concebir como un cubo de tres o cuatro dimensiones. Los usuarios pueden acceder a una porción de la base de datos a lo largo de cualquiera de sus dimensiones. Dado que es muy común representar a un modelo dimensional como una tabla de hechos rodeada por las tablas de dimensiones, frecuentemente se le menciona también como modelo estrella o se denomina Snowflake o copo de nieve a ramificaciones a partir de las tablas de dimensiones.

### 2.4.6 DEFINICIÓN DE ETL

Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, DataMart, o DataWarehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

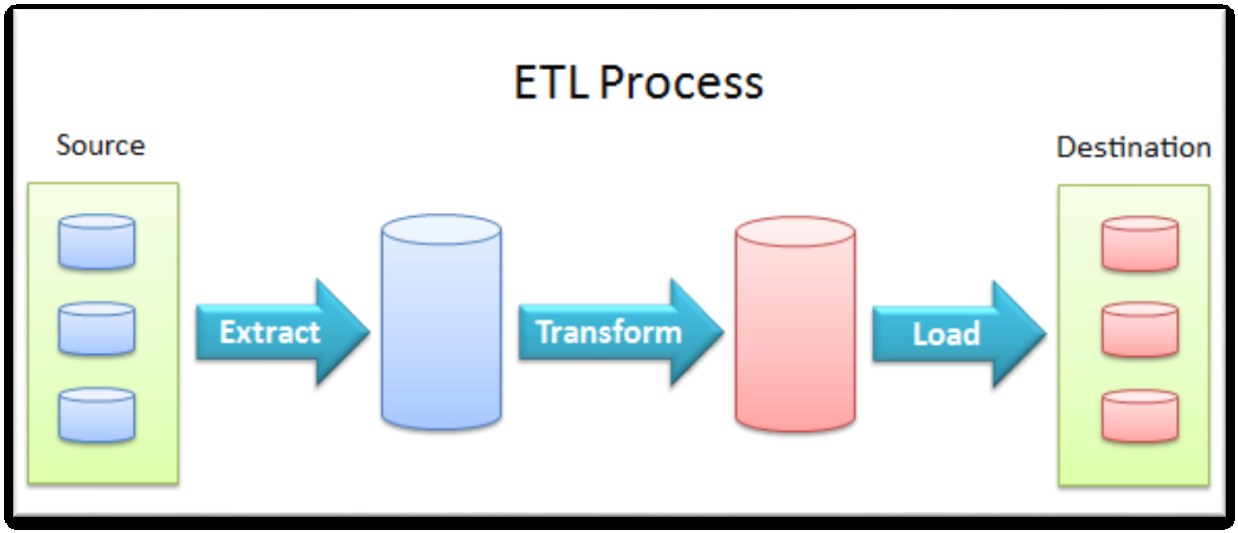
En ingles su acrónimo es *Extract, Transform and Load* («extraer, transformar y cargar», frecuentemente abreviado ETL[[9]](#footnote-9)). Los procesos ETL[[10]](#footnote-10) también se pueden utilizar para la integración con sistemas heredados. Se convirtieron en un concepto popular en los años

1970.

2.4.6.1 HERRAMIENTAS

Las herramientas de ETL se utilizan para extraer datos de fuentes de datos homogéneas o heterogéneas y de esa forma transformar los datos para almacenarlos en el formato o estructura adecuados para fines de consulta y análisis. Paso posterior cargarlo en la base de datos. En el blog de tecnología Techroba[[11]](#footnote-11) publicado en diciembre de 2015 se dice que usualmente en las herramientas ETL, las tres fases se ejecutan en paralelo ya que la extracción de datos lleva tiempo, así que mientras se extraen los datos se ejecuta otro proceso de transformación, procesando los datos ya recibidos y preparando los datos para cargar y tan pronto como haya datos listo para ser cargado en el objetivo, la carga de datos comienza sin esperar la finalización de las fases anteriores.

Imagen 13. Proceso ETL



Fuente: Adaptado de Data Science (s. f.) Recuperado el 24 de febrero de 2018, a partir de https://www.datasciencecentral.com/profiles/blogs/10-open-source-etl-tools

2.4.6.1.1 HERRAMIENTAS ETL DE CÓDIGO ABIERTO

A continuación se listas las más conocidas de código abierto:

* TALEND

Proporciona múltiples soluciones para la integración de datos, tanto de código abierto como de edición comercial. Ofrece una interfaz basada en Eclipse, flujo de diseño de arrastrar y soltar y amplia conectividad con más de 400 conectores de aplicaciones pre configurados para unir bases de datos, mainframes, sistemas de archivos, servicios web, aplicaciones empresariales empaquetadas, aplicaciones de Softwareas-a-Service, basadas en la nube y más.

* SCRIPTELLA

Es una herramienta de ejecución de scripts escrita en Java. Su enfoque principal es la simplicidad. No tiene que estudiar otro lenguaje complejo basado en XML, se puede usar SQL u otro lenguaje de scripting adecuado para realizar las transformaciones requeridas. Scriptella tiene licencia bajo la Licencia Apache, Versión 2.0

* KETL

Es una herramienta de primer nivel. La plataforma de integración de datos está construida con una arquitectura portátil basada en Java y una configuración abierta basada en XML y lenguaje de trabajo. Las características de KETL compiten exitosamente con los principales productos comerciales disponibles en la actualidad.

Los puntos destacados incluyen:

* Soporte para la integración de herramientas de seguridad y gestión de datos.
* Escalabilidad comprobada en múltiples servidores y CPU y cualquier volumen de datos.
* Sin necesidad adicional de herramientas de programación, dependencia y notificación de terceros.

* INTEGRADOR DE DATOS PENTAHO (KETTLE)

La integración de datos Pentaho es una aplicación y biblioteca Java (Swing). Kettle es un intérprete de procedimientos escritos en formato XML. Las características y los componentes son un poco menos compresivos que los de Talend, sin embargo, esto no restringe la complejidad de los procedimientos de ETL que se pueden implementar. Kettle proporciona un motor de JavaScript (así como uno de Java) para afinar el proceso de manipulación de datos.

Kettle es también una buena herramienta, con todo lo necesario para crear incluso procedimientos complejos de ETL. Kettle (PDI) es la herramienta predeterminada en Pentaho Business Intelligence Suite. Los procedimientos también se pueden ejecutar fuera de la plataforma Pentaho, siempre que estén instalados todas las bibliotecas de Kettle y el intérprete de Java.

* JEDOX

Jedox es una solución de código abierto de BI para la gestión del rendimiento, que incluye planificación, análisis, informes y ETL. El núcleo abierto consta de un servidor OLAP en la memoria, el servidor ETL y las bibliotecas del cliente OLAP.

Potentemente compatible con el servidor Jedox OLAP como un sistema de origen y destino, Jedox ETL está específicamente diseñado para enfrentar los desafíos del análisis OLAP. Trabajar con cubos y dimensiones no podría ser más fácil. Genere con flexibilidad jerarquías de tiempo que se necesitan con frecuencia y transforme de manera eficiente el modelo relacional de sistemas fuente en un modelo OLAP - con

JEDOX ETL.

2.4.6.1.2 HERRAMIENTAS ETL DE CÓDIGO PRIVATIVO

A continuación se listas las más conocidas de código abierto[[12]](#footnote-12):

 SQL SERVER

Es un sistema para la gestión de bases de datos producido por Microsoft basado en el modelo relacional. Sus lenguajes para consultas son T-SQL y ANSI SQL. Sus principales características son:

* Rendimiento ultra rápido
* Descubrimiento de datos rápido
* Reporting
* Análisis predictivo

* DATA WAREHOUSING SAP

Nuestra plataforma de almacenamiento de datos permite capturar, almacenar, analizar y gestionar datos en un DataWarehouse seguro y escalable, para luego recuperarlos y poder convertirlos en información lista para tomar decisiones. Sus principales características son:

* + Explorar y analizar grandes volúmenes de datos en tiempo real.
  + Observar tendencias, patrones y relaciones que de otro modo no estarían a la

vista.

* + Brindar a los usuarios acceso rápido a información pertinente en cualquier lugar de la empresa.

* DB2

Ventajas y Características

* + Autonómico
  + Soporte NoSQL
  + Seguridad
  + Aceleración Blu
  + Análisis de DataWarehouse
  + Rendimiento Optimizado
  + Compatibilidad SQL
  + Almacenaje Optimizado
  + Alta Disponibilidad

* PENTAHO

Acelera la realización de grandes volúmenes de datos con la solución más completa para el análisis de datos grandes. Sus principales características son:

* + - Soporte No SQL
    - Acceso, exploración y organización de fuentes de datos
    - Capacidades de Normalización
    - Conectividad a Cualquier tipo de datos
    - Arrastra y soltar diseño sin generación de código
    - Arquitectura Moderna

* ORACLE WAREHOUSE BUILDER

Es una solución de integración de datos de almacenamiento. Sus principales

características son:

* + - Permite cualquier cliente Oracle, escalable.
    - Incluye repositorio de metadatos, soporte de administración.

## 2.5 MINERÍA DE DATOS

### 2.5.1 OBJETIVO DE MINERÍA DE DATOS

Seleccionar las variables de los empleados para poderlos agrupar y asociar, de esa forma detectar posibles tendencias o patrones de comportamiento relacionadas con el desempeño laboral de los empleados.

Una de las ventajas que comenta Gomez-Flechoso, A.J. (1998) de contar con bases de datos es la posibilidad de descubrir información de interés y adquirir conocimiento mediante el análisis de los datos. Sin embargo, el volumen mismo de las bases de datos es con frecuencia una limitante para análisis manuales por lo que se han desarrollado tecnologías especializadas que faciliten su manejo.

La minería de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Algunas de las definiciones que han dado diferentes autores para la minería de datos son:

* Witten y Frank (2005), la minería de datos es el proceso de descubrir patrones en los datos. Los datos se presentan en grandes cantidades. Los patrones descubiertos deben ser significativos de manera que se permitan ventajas, por lo general, de tipo económicas.

* Sumathi y Sivanandam (2006), la minería de datos es el proceso eficiente, no trivial, de extraer información valiosa (patrones y tendencias) de una gran colección de datos.

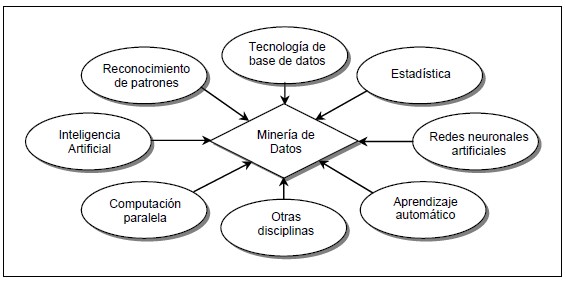
* Hand et al. (2001), la minería de datos es el análisis de conjuntos de datos observados, a menudo extensos, para encontrar relaciones insospechadas y resumir los datos en forma comprensible y útil para el usuario de la información.

* Berry y Linoff (2004), la minería de datos, es la exploración y análisis de grandes cantidades de datos para descubrir reglas y patrones significativos.

Los autores encuentran en estas definiciones datos en común en la extracción de información útil, estos mismos son patrones, asociaciones y relaciones entre los datos para los usuarios finales. Debido a las muchas formas de definir la minería de datos, Hernández et al. (2004) resume el objetivo de la minería de datos como el proceso de convertir datos extraídos de grandes volúmenes de datos en conocimiento útil. Para Witten y Frank (2005), el establecimiento de patrones es una práctica útil, puesto que permite explicar ―algo‖ con relación a los datos que se tienen.

De la misma manera, autores como Vazirgiannis et al. (2003) y Geng y Hamilton (2007) consideran que la minería de datos facilita el relevar patrones existentes en el conjunto de datos, los cuales generan información válida, novedosa, útil, comprensible y explotable. Estos factores han llevado a considerar a la minería de datos como un campo interdisciplinario, el cual es desarrollado de manera conjunta o como apoyo de otras disciplinas como: tecnología de bases de datos, sistemas para la toma de decisiones, estadística, inteligencia artificial, aprendizaje automático, redes neuronales, reconocimiento de patrones, visualización de datos, computación paralela, entre otras.

Imagen 14. Asociación diversas disciplinas de minería de datos.



Fuente: Adaptado de Hernández et al. (2004)

Una aplicación de minería de datos podría identificar varios grupos en los datos que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones.

Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos, pero pertenecen a todo el proceso

KDD (Knowledge Discovery in Databases) como pasos adicionales.

Los términos relacionados con el dragado de datos, la pesca de datos y el espionaje de los datos se refieren a la utilización de métodos de minería de datos a las partes de la muestra que son (o pueden ser) demasiado pequeños para las inferencias estadísticas fiables que se hicieron acerca de la validez de cualquiera de los patrones descubiertos. Estos métodos pueden, sin embargo, ser utilizados en la creación de nuevas hipótesis que se prueban contra las poblaciones de datos más grandes. Hernández (2004) define que un proceso típico de minería de datos consta de los siguientes pasos generales:

* Selección del conjunto de datos:

Tanto en lo que se refiere a las variables objetivo (aquellas que se quieren predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.

* Análisis de las propiedades de los datos:

En especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

* Transformación del conjunto de datos de entrada:

Se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema. A este paso también se le conoce como pre-procesamiento de los datos.

* Seleccionar y aplicar la técnica de minería de datos:

Se construye el modelo predictivo, de clasificación o segmentación.

* Extracción de conocimiento:

Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesamiento diferente de los datos.

* Interpretación y evaluación de datos:

Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Si el modelo final no superara esta evaluación, el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido. Una vez validado el modelo, éste ya está listo para su explotación.

Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del DataMining Group, que está estandarizando el lenguaje PMML [[13]](#footnote-13) (*Predictive Model Markup Language*), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

Tradicionalmente las técnicas de minería de datos se aplicaban sobre información contenida en almacenes o bodegas de datos tal como lo dice Hernández Orallo, J. (2004).

De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos no estructurados como información contenida en archivos de texto, en Internet, etc.

### 2.5.2 TIPOS DE MINERÍA DE DATOS

La minería de datos es un paso esencial de un proceso más amplio cuyo objetivo es el descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD). Fayyad et al. (1996). La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc. Los más importantes de ellos son:

* La importancia que han cobrado los datos no estructurados (texto, páginas de

Internet, etc.).

* La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.

* La exigencia de que los procesos funcionen prácticamente en línea (por ejemplo, en casos de fraude con una tarjeta de crédito).

* Los tiempos de respuesta. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente pues esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real.

2.5.2.1 PREDICCIÓN

Muchas formas de minería de datos son predictivos. Por ejemplo, un modelo podría predecir el ingreso basado en la educación y otros factores demográficos. Las predicciones tienen una probabilidad asociada y las probabilidades de predicción son también conocidas como confianza. Algunas formas de minería de datos predictiva generan reglas, las cuales son condiciones que implican una salida dada. Por ejemplo, una regla podría especificar que una persona que tiene un grado universitario y vive en cierta colonia probablemente tiene un ingreso mayor que el promedio en la región.

2.5.2.2 AGRUPACIÓN

La agrupación es otra forma en la que la minería de datos identifica grupos naturales en los datos. Por ejemplo, un modelo podría identificar el segmento de la población que tiene un ingreso dentro de un rango específico, que tiene un buen registro de manejo, y que arrienda un carro nuevo con base anual. Muñoz y Rodrigo (2002) señalan que el agrupamiento ha sido reconocido como un método eficaz para tratar problemas de clasificación de variables de estudio para la búsqueda de patrones y tendencias.

En el agrupamiento existen dos tipos principales (Larose, 2005): el agrupamiento jerárquico que se caracteriza por el desarrollo recursivo de una estructura en forma de árbol, y el agrupamiento particional que organiza los registros dentro de *k* grupos. Los métodos particionales tienen ventajas en aplicaciones que involucran gran cantidad de datos y los jerárquicos por lo general se utilizan para identificar el número deseado de grupos.

### 2.5.3 FUNCIONES Y TAREAS DE LA MINERÍA DE DATOS

La Minería de Datos se apoya en la aplicación de métodos matemáticos de análisis, utilizando diferentes algoritmos y técnicas de clasificación, tales como clustering, regresión, inteligencia artificial, redes neuronales, reglas de asociación, árboles de decisión, algoritmos genéticos, entre otras, que son de gran utilidad para llevar a cabo el análisis inteligente de grandes volúmenes de información digital. La minería de datos relacionada con la educación se denomina ―Minería de datos educativa‖ como lo define Alaa el-Halees (2009).

Estos elementos permiten que los algoritmos aprendan conceptos por sí solos, sin tener que ser programados. Es decir, se trata de un conjunto de reglas abstractas que por sí solas son construidas, lo que se ha conseguido es una ―autoconfiguración‖ propiamente dicha.

 Problemas de clasificación:

Cuando la variable a predecir es un conjunto de estados discretos o categóricos de los cuales a su vez pueden ser:

* Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}, etc.
* Múltiple: Comprará {Producto1, Producto2…}, etc.
* Ordenada: Riesgo {Bajo, Medio, Alto}.

Al igual que la clasificación, la regresión es una tarea de aprendizaje inductivo que ha sido ampliamente estudiada y utilizada. Los modelos de regresión predicen variables y valores numéricos por ejemplo las ventas de una empresa a partir de los precios a fijar.

Estos métodos proporcionan una forma directa de descubrir y representar la similitud en los datos, formando conjuntos de datos similares entre ellos. Una vez que se crean los grupos, los algoritmos pueden predecir a qué conjunto corresponde una nueva instancia, lo que les permite implementar análisis predictivos.

Imagen 15. Tipología de los algoritmos de minería de datos



Fuente: Adaptado de Gironés, Jordi, et al. Minería de datos: modelos y algoritmos,

Editorial UOC, 2017. ProQuest Ebook Central,

http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=5045398

2.5.3.1 MINERÍA DE DATOS SUPERVISADA

En el libro *Minería de datos: modelos y algoritmos* del autor Gironés, Jordi menciona que ―Los métodos supervisados son algoritmos que basan su proceso de aprendizaje en un juego de datos de entrenamiento convenientemente etiquetado. Por etiquetado entendemos que para cada ocurrencia del juego de datos de entrenamiento conocemos el valor de su atributo objetivo o clase. Esto le permitirá al algoritmo poder deducir una función capaz de predecir el atributo objetivo para un juego de datos nuevo. Las grandes familias de algoritmos de aprendizaje supervisado son: Algoritmos de clasificación y

Algoritmos de regresión.‖ (Gironés, Casas, & Minguillón, 2017).

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento que consisten en pares de objetos normalmente vectores; datos de entrada y los resultados deseados. La precisión de la función aprendida depende en gran medida de cómo el objeto de entrada está representado. La salida de la función puede ser un valor numérico como en los problemas de regresión o una etiqueta de clase como en los de clasificación. El objetivo de este aprendizaje es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber ajustado los parámetros y haber visto una serie de ejemplos con los datos de entrenamiento.

El desempeño del algoritmo se puede medir utilizando un conjunto de pruebas independiente del de entrenamiento que se usa con el mismo. Los algoritmos más utilizados son las redes neuronales, las máquinas de vectores de soporte; el algoritmo de los K-vecinos más cercanos *(KNN)*, los modelos de mixturas; el clasificador bayesiano ingenuo, los árboles de decisión y las funciones de base radial.

2.5.3.1.1 KNN VECINOS MÁS CERCANOS

Es un algoritmo de aprendizaje supervisado del tipo *clasificación*, con los parámetros y los datos de entrenamiento su objetivo será clasificar correctamente todas las instancias nuevas. En contraste con otros algoritmos de aprendizaje supervisado, el algoritmo KNN no genera un modelo fruto del aprendizaje con datos de entrenamiento, sino que el aprendizaje sucede en el mismo momento en el que se pide clasificar una nueva instancia.

A este tipo de algoritmos se les llama métodos de aprendizaje perezoso o *Lazy learning method* en inglés. El funcionamiento del algoritmo es muy simple. Para cada nueva instancia a clasificar, se calcula la distancia con todas las instancias de entrenamiento y se seleccionan las k instancias más cercanas.

La clase de la nueva instancia se determina como la clase mayoritaria de sus k instancias más cercanas afirman Wagsta, K., Cardie, C.

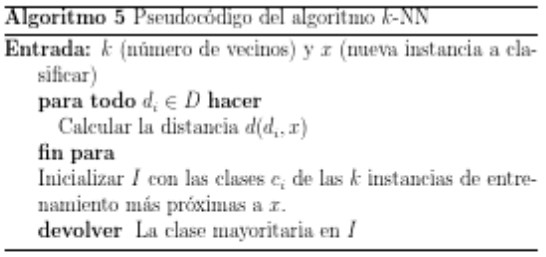
El funcionamiento del método se detalla en la imagen 22 y se describe a continuación:

1. Fijamos un valor para k, habitualmente pequeño.

1. Dada una nueva instancia x del juego de datos de prueba, el algoritmo selecciona las k instancias del conjunto de datos de entrenamiento más cercanas, de acuerdo con la métrica de similitud utilizada.

1. Finalmente, se asigna la instancia x a la clase más frecuente de entre las k instancias seleccionadas como más cercanas.

Imagen 16. Pseudocódigo del algoritmo KNN

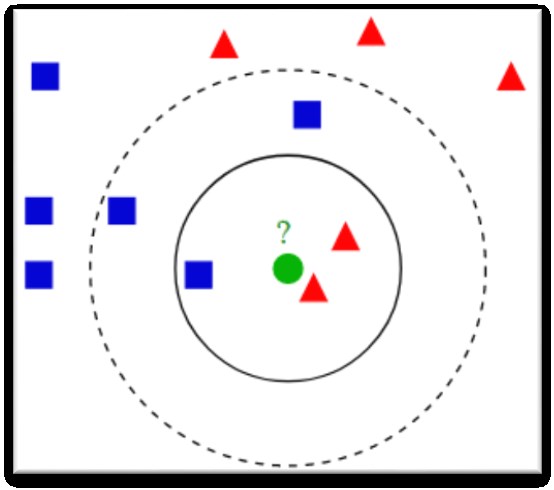


Fuente: Adaptado de algoritmo KNN. Recuperado 12 de febrero de 2018 del sitio web

https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf

A partir de un set de datos entrenados y clasificados se procede a la asignación de la clasificación para un segundo conjunto de datos de prueba.

Imagen 17. Búsqueda de clases a la que pertenece la clasificación del punto.



Fuente: Elaboración propia

El algoritmo se basa en la idea que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del set de entrenamiento más próximo.

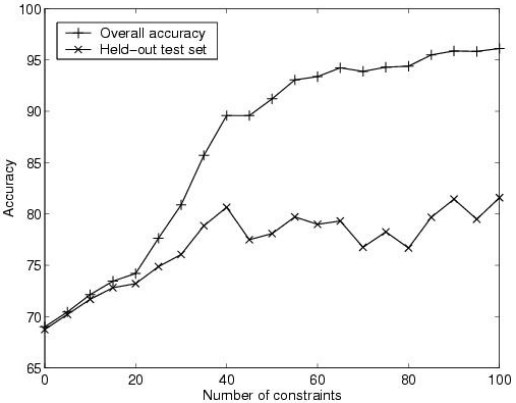
2.5.3.2.1 SVN MAQUINAS DE SOPORTE VECTORIAL

Las máquinas de soporte vectorial en inglés *(Support Vector Machines o SVM)* son algoritmos de aprendizaje supervisado capaz de resolver problemas de clasificación, tanto lineales como no lineales. Actualmente es considerado uno de los algoritmos más potentes en reconocimiento de patrones. Su eficiencia y los buenos resultados obtenidos en comparación con otros algoritmos han convertido esta técnica en la más utilizada en campos como el reconocimiento de textos y habla, predicción de series temporales y estudios sobre bases de datos de marketing, entre otros, pero también en otros campos como la secuenciación de proteínas y el diagnóstico de varios tipos de cáncer.

Las SVM utilizan las técnicas de optimización cuadrática propuestas por el matemático italiano Giuseppe Luigi Lagrange para buscar maximizar el margen entre los puntos pertenecientes a los distintos grupos a clasificar, de tal forma que el máximo número de futuros puntos queden bien clasificados. El objetivo de las SVM es encontrar el hiperplano óptimo que maximiza el margen entre clases del conjunto del set de datos de entrenamiento.

La muestra un gráfico de dispersión de un ejemplo del set de datos de entrenamiento, donde se ve las dos clases presentes en el conjunto de datos. Se observar que a partir de un hiperplano se puede separar las dos clases de datos de forma clara. Hay que tener en cuenta que el hecho de disponer de un hiperplano separador no garantiza, en absoluto, que este sea el mejor de todos los posibles hiperplanos separadores, ya que puede configurarse para tener infinitos hiperplanos.

Imagen 18. Concepto de margen e hiperplano



Fuente: Adaptado de hiperplano (s. f.). Recuperado 18 de febrero de 2018 del sitio web

https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf

De este modo, cuanto más cerca este el hiperplano de los puntos, menor será su margen y el objetivo es encontrar el mejor separador óptimo que maximice el margen del conjunto y set de datos de entrenamiento.

2.5.3.2 MINERÍA DE DATOS NO SUPERVISADA

Dentro de la minería de datos las técnicas y métodos que usan los algoritmos no supervisados son utilizadas para la detección de patrones ocultos en la información y fuentes de datos de la empresa. Dichos patrones representan por sí mismos información útil que puede ser utilizada directamente en la toma de decisiones. Así mismo, Palma, Méndez, José Tomás, and Morales, Roque Marín dicen que este tipo de métodos no supervisados se pueden clasificar en tres grandes grupos, dependiendo del esquema de organización y como se quiera plantear la actualizar las características y de los datos:

* Métodos de aprendizaje basados en la regla de Hebb. Una de las primeras aproximaciones al aprendizaje no supervisado fue la realizada por Hebb que ha supuesto una de las aportaciones fundamentales en el campo de las redes de neuronas. Basados en ella, se han propuesto diversos tipos de reglas hebbianas empleadas con diversos fines, aunque una de las aplicaciones más importantes es en el análisis de componentes principales.

* Métodos de aprendizaje competitivos. En este tipo de sistemas el objetivo es agrupar o categorizar los datos que sean similares. Generalmente, se trata de similitud geométrica en el espacio de entrada. Estos modelos utilizan redes neuronales que compiten para ser activadas.

A diferencia de los métodos anteriores, donde varias neuronas pueden ser activadas al mismo tiempo, en este tipo de sistemas sólo una de ellas será la ganadora. Por tanto, la red tratará de agrupar los patrones similares que serán representados por una neurona de la red.

• Métodos de aprendizaje basados en modelos de teoría de la información. En 1988 Linsker (Linsker, 1988) propuso por primera vez un esquema de aprendizaje basado en los principios de la teoría de la información. Este tipo de sistemas trata de maximizar la cantidad de información que se conserva en el procesamiento de los datos minimizando la entropía. Este tipo de métodos tiene multitud de aplicaciones prácticas, entre ellas, el análisis de componentes principales y sistemas de compresión de información.

La mayoría de los métodos descritos antes tienen como objetivo seleccionar las mejores reglas de asociación teniendo en cuenta los factores de soporte y de confianza para obtener patrones consistentes e interesantes. Las reglas de asociación que define el autor

Agrawal et al.(1993b) dieron comienzo al algoritmo Apriori que describe Agrawal, R.,

Srikant (1994) con el objetivo principal de proporcionar un número reducido de reglas para seleccionar las más útiles para los propósitos de la clasificación de atributos y factores interesantes en los patrones de conocimiento.

2.5.3.2.1 CLUSTERING

Este algoritmo intenta encontrar patrones en los datos identificando principalmente los que se forman como grupos separados. Esos criterios son por lo general *distancia* o *similitud*. Cada grupo encontrado se puede usar para realizar una segmentación de clientes/empleados y de esa forma ofrecer así distintos productos a cada grupo. El objetivo principal es encontrar grupos de instancias que en un cluster sean similares entre sí y diferentes de las otras en otros clusters. Otra posible aplicación es agrupar documentos por temática, donde cada cluster o grupo pertenece a un tipo de documento, sin embargo puede usarse como paso previo a otras técnicas de aprendizaje. Algunos ejemplos son exploración de datos y pre-procesamiento de datos.

2.5.3.2.2 K-MEANS

El algoritmo K-medias (o K-Means) propuesta por J. B. MacQueen en 1967 (Berry y Linoff, 2004) es probablemente el algoritmo de agrupamiento más conocido. El algoritmo está basado en la minimización de la distancia interna (la suma de las distancias de los patrones asignados a un agrupamiento al centroide de dicho agrupamiento). De hecho, este algoritmo minimiza la suma de las distancias al cuadrado de cada patrón al centroide de su agrupamiento.

El algoritmo es sencillo y eficiente. Además, procesa los patrones secuencialmente (por lo que requiere un almacenamiento mínimo).

Sin embargo, está sesgado por el orden de presentación de los patrones (los primeros patrones determinan la configuración inicial de los agrupamientos) y su comportamiento depende enormemente del parámetro K.

El algoritmo de K-medias, también llamado algoritmo de Lloyd en la comunidad informática, es un método de clustering o agrupamiento que consiste en clasificar un conjunto de elementos con base en propiedades de estos de forma iterativa. El algoritmo fue propuesto por primera vez por Stuart Lloyd en 1957 como una técnica para modulación por impulsos codificados. Mediante el método de K-medias se espera obtener resultados que revelen patrones del conjunto de datos de entrada. La técnica de agrupamiento se fija en el número *k* de clusters al inicio de la ejecución y se asignan elementos al cluster en función de la distancia, empleando para ello funciones como la distancia euclídea.

Otra característica importante es que emplea la media estadística para el cálculo los nuevos clusters. Concretamente, se calcula del cluster su centroide; esto es el punto resultante de la media de todos los elementos asignados al cluster en cuestión. Los dos pasos principales del procedimiento son:

1. Paso de asignación:

Se asigna a cada elemento el cluster o grupo más cercano, aplicando una función para el cálculo de distancia.

1. Paso de actualización:

Consistente en calcular los nuevos centroides para cada clúster. Aunque el algoritmo presenta una complejidad computacional difícil, existen heurísticos que hacen que converja rápidamente a un óptimo local.

También se requiere de un método de inicialización usando dos principalmente:

* Método Forgy:

Consistente en la toma de k observaciones de forma aleatoria como centros del cluster y luego empezar con el paso de asignación.

* Método Partición Aleatoria:

En el que se asigna aleatoriamente un cluster a cada observación y se sigue con el paso de actualización. Estos métodos funcionan mejor respecto de otros métodos heurísticos que se empleen. K-medias converge cuando al realizar el paso de asignación ningún elemento cambia de cluster.

El procedimiento que sigue la técnica de K-medias es el siguiente (Jain et al., 1999;

Larose, 2005):

Imagen 19. Ecuación de Centroide



Fuente: Adaptado de Ecuación del centroide (s. f.). centroide.png (156×50). Recuperado

24 de marzo de 2018, del sitio web [http://scikit-](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

[learn.org/stable/\_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

1. Selección aleatoria de *K* puntos o elementos, haciendo que éstos representen los

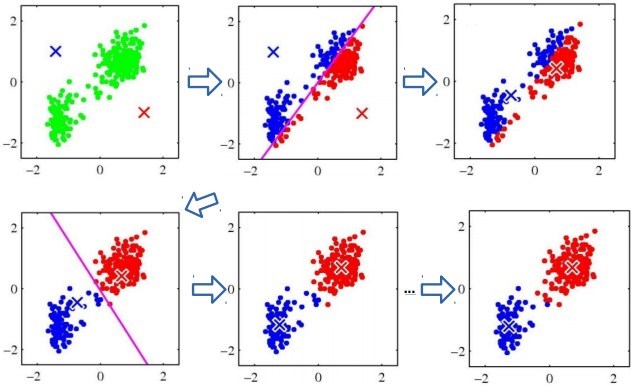
―centros‖ de los grupos.

1. Asignación de cada uno de los elementos restantes al centro más cercano. Esto es, la distancia mínima entre el elemento y el centro. Por lo general, la medida de distancia utilizada es la euclidiana.

1. Una vez que todos los elementos han sido asignados, se recalculan los K centros.

Y por último se repiten los pasos 2 y 3 hasta que los centros ya no se modifiquen.

Imagen 20. Asignación de los centroides



Fuente: CM Bishop, Pattern Recognition and Machine Learning, Fig 9.1, p426

Para asignar los registros a los grupos cuyo centro es el más cercano, se utiliza la distancia euclidiana cuadrática definida como Clementine (2006a):

Imagen 21. Ecuación de Centroide



Fuente: Adaptado de Ecuación del centroide (s. f.). centroide.png (156×50). Recuperado

24 de marzo de 2018, del sitio web [http://scikit-](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

[learn.org/stable/\_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png) Dónde:

* *X i* vector de los campos de entrada para el registro i
* *C j* centro del grupo para la región j
* *Q* número de campos de entrada
* *x qi* valor del q-ésimo campo de entrada para el i-ésimo registro
* *c qj* valor del q-ésimo campo de entrada para el j-ésimo registro

Para actualizar el valor de los centros en los grupos, éstos se calculan como el vector medio de los registros establecidos en dicho grupo, *C j = Xj*, donde los campos del vector medio *Xj* son calculados de acuerdo con la siguiente ecuación:

Imagen 22. Ecuación de Centroide



Fuente: Adaptado de Ecuación del centroide (s. f.). centroide.png (156×50). Recuperado

24 de marzo de 2018, del sitio web [http://scikit-](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

[learn.org/stable/\_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

Donde *nj* es el número de registros en el grupo *j,* y *xqi* es el q-ésimo valor para el registro *i* que es asignado al grupo *j*.

A pesar de ser antiguo y falto de robustez, este algoritmo es de uso muy extendido, en áreas como segmentación de mercados, visión para computadoras, geo estadística, astronomía o recuperación de la información. Es empleado además como pre

procesamiento para otros algoritmos, por ejemplo para búsqueda de configuración inicial.

2.5.3.3 HÍBRIDOS

Este tipo de algoritmos combinan los dos algoritmos anteriores para poder clasificar de manera adecuada. Se tiene en cuenta los datos marcados y los no marcados. A continuación se detallan los más importantes métodos.

2.5.3.3.1 ALGORITMO DE ÁRBOLES DE DECISIÓN DE SQL

(MICROSOFT)

Es un algoritmo de clasificación y regresión usado para el modelo de predicción de atributos discretos y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

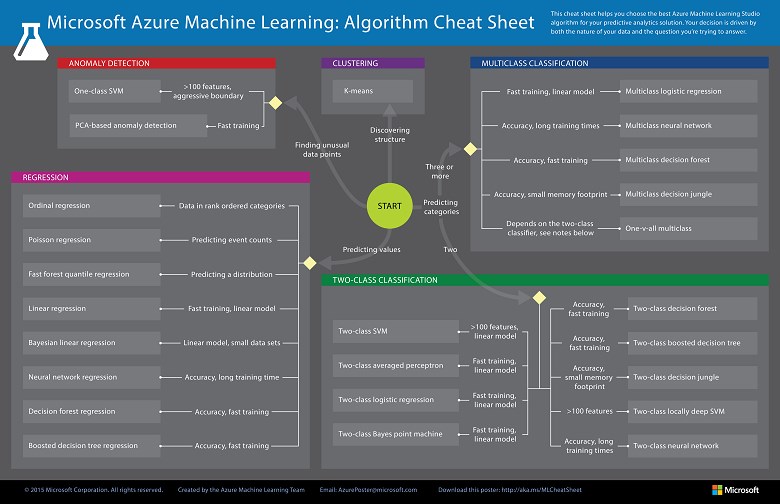
Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión[[14]](#footnote-14).

Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

2.5.3.3.2 ALGORITMO DE CLÚSTERES DE SQL (MICROSOFT)

Es un algoritmo de segmentación suministrado por Analysis Services. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual.

Imagen 23. Microsoft-machine-learning-algorithm-cheat-sheet-v6



Fuente: Adaptado de Pakalra. (s. f.). Algoritmos de aprendizaje automático. Recuperado 24 de junio de 2018, a partir de [https://docs.microsoft.com/es-es/azure/machine-](https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-cheat-sheet)

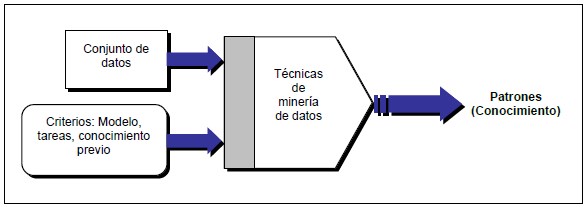
[learning/studio/algorithm-cheat-sheet.](https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-cheat-sheet)

### 2.5.3 TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de Machine Learning nos abren una amplia ventana al mundo de los datos, ofreciéndonos grandes posibilidades para mejorar la eficiencia y el rendimiento de nuestro negocio y elevar, en consecuencia, la competitividad de nuestra empresa. En este sentido, apenas hemos comenzado a vislumbrar el gran potencial que nos ofrece el aprendizaje automático para abordar problemas reales y crear modelos que proporcionen predicciones de alto valor, permitiendo, de esta forma, tomar decisiones más adecuadas y desarrollar mejores acciones de negocio.

Imagen 24. Proceso de extracción de patrones mediante el uso de técnicas de minería de

datos.



Fuente: Adaptado de Hernández et al. (2004)

Por tanto, es conveniente aprovechar la oportunidad de obtener información valiosa a partir de nuestros datos y no perder de vista las ventajas competitivas que nos puede brindar el asomarnos al extenso e incipiente universo del Machine Learning. La técnica más utilizada en minería de datos es la de *clasificación* que emplea métodos como *árboles de decisión* o *redes neuronales*. Cada proceso de clasificación que se realiza implica un aprendizaje y una propia clasificación. Tomado de Larose D. (2005). Ese aprendizaje es donde se entrenan los datos mediante los diferentes algoritmos, para posteriormente realizar las pruebas y comprobar resultados. En esta etapa del proyecto seleccionamos los algoritmos posibles que nos ayudarán a determinar los factores que afectan el aprovechamiento académico.

2.5.3.1 REDES NEURONALES

Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

* El Perceptrón.
* El Perceptrón Multicapa.
* Los mapas auto organizados, conocidos como redes de Kohonen.

Una red neuronal es básicamente una interconexión de neuronas que trabajan entre sí para producir una salida, en la cual se generan procesos necesarios asociados al aprendizaje como respuesta a un estímulo generado en el ambiente. Haykin citado en el libro de Discovering Knowledge in Data: An Introduction to Data Mining del autor

Larose D.

Propone la siguiente definición: ―Una red neuronal es un procesador masivamente paralelo distribuido que es propenso, por naturaleza, a almacenar conocimiento experimental y hacerlo disponible para su uso‖. Con la ayuda de las redes neuronales se puede llegar a identificar factores en los alumnos con buenas o malas perspectivas de aprovechamiento académico. Calcular la probabilidad de que un alumno pueda desertar.

Clasificar los diferentes atributos de los alumnos y explorar los factores relacionados.

2.5.3.2 ALGORITMO DE RED NEURONAL DE SQL (MICROSOFT)

En *SQL Server Analysis Services* el algoritmo de red neuronal combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada.

Los modelos de minería de datos construidos con este algoritmo pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o solo para la predicción. El número de redes que contiene un único modelo de minería de datos depende del número de estados que contienen las columnas de entrada y las columnas de predicción que utiliza el modelo.

Las redes artificiales de neuronas tratan, en cierto modo, de replicar el comportamiento del cerebro, donde tenemos millones de neuronas que se interconectan en red para enviarse mensajes unas a otras. Esta réplica del funcionamiento del cerebro humano es uno de los ―modelos de moda‖ por las habilidades cognitivas de razonamiento que adquieren.

El reconocimiento de imágenes o vídeos, por ejemplo, es un mecanismo complejo para mostrar una red neuronal para hacer señala Simon Haykin (1999).

2.5.3.3 AGRUPAMIENTO O CLUSTERING

Un algoritmo de agrupamiento *(en inglés, clustering)* es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

Berry y Linoff (2004) y Sumathi y Sivanandam (2006) consideran al agrupamiento como la tarea de dividir una población heterogénea en un número de subgrupos homogéneos de acuerdo a las similitudes de sus registros. Dentro de esta tarea existen dos tipos principales de agrupamiento (Larose D, 2005): el jerárquico que se caracteriza por el desarrollo recursivo de una estructura en forma de árbol, y el particional que organiza los registros dentro de k grupos.

Una característica de este tipo de agrupamiento es el establecer a priori el número de grupos de entrada (*k*), por lo que en la práctica es necesario repetir la prueba estableciendo diferentes números de grupos, eligiendo la solución que mejor se adapte al objetivo del problema (Jain et al., 1999). Un método sugerido por Milligan (1980, 1985) y Hair et al. (1995) para determinar el número de grupos de entrada (*k*) es usar el resultado obtenido por algún algoritmo jerárquico, mediante el cual se obtiene el número deseado de grupos, posteriormente se aplica algún algoritmo particional.

2.5.3.4 CLASIFICACIÓN

Es una de las principales tareas en el proceso de minería de datos que se emplea para asignar datos a un conjunto predefinido de variables (Vazirgiannis et al., 2003). El objetivo de la clasificación es encontrar algún tipo de relación entre los atributos de entrada y los registros de salida para comprender el comportamiento de los datos, así mediante el conocimiento extraído se puede predecir el valor de un registro desconocido (Sumathi y Sivanandam, 2006). Sin embargo, el mayor problema de la clasificación es que muchas veces no es representativo y no proporciona un conocimiento detallado, sólo otorga predicciones.

Algunos algoritmos comprendidos en esta tarea son: clasificación bayesiana, árboles de decisión, redes neuronales artificiales, entre otros (Hand et al., 2001; Vazirgiannis et al.,

2003).

2.5.3.5 REGRESIÓN

La regresión es el aprendizaje de una función cuyo objetivo es predecir valores de una variable continua a partir de la evolución de otra variable también continua, la cual por lo general es el tiempo (Sumathi y Sivanandam, 2006). En la regresión, la información de salida es un valor numérico continuo o un vector con valores no discretos (Vazirgiannis et al., 2003)*.* Ésta es la principal diferencia respecto a la clasificación donde el valor a predecir es numérico.

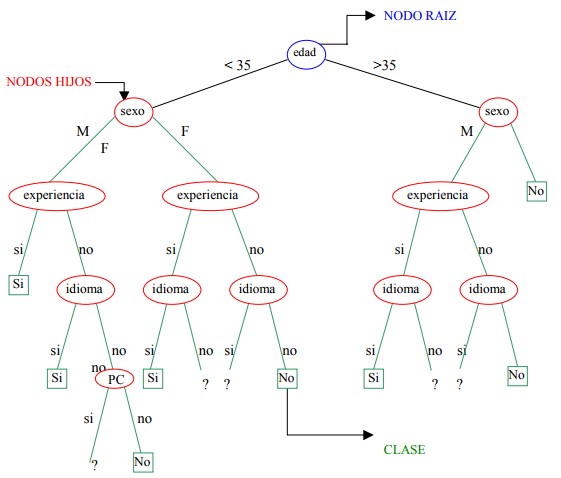
Si sólo se dispone de una variable definida se trata de un problema de regresión simple, mientras que si se dispone de varias variables se trata de un problema de regresión múltiple. A esta tarea también se le conoce como: interpolación, cuando el valor o valores predichos están en medio de otros; o estimación, cuando se predice valores futuros Hernández et al., (2004).

2.5.3.6 ÁRBOLES DE DECISIÓN

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Aparte, como dicen Zhu, X. y I. Davidson. (2007) los arboles de decisión son una técnica de minería de datos que establece un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Se puede decir que los arboles de decisión se adecuan más a la clasificación para poder determinar las clases que se puedan generar, y por tal motivo poder identificar a que clase pertenece un objeto.

Imagen 25. Árbol de decisión, ejemplo entrevista laboral

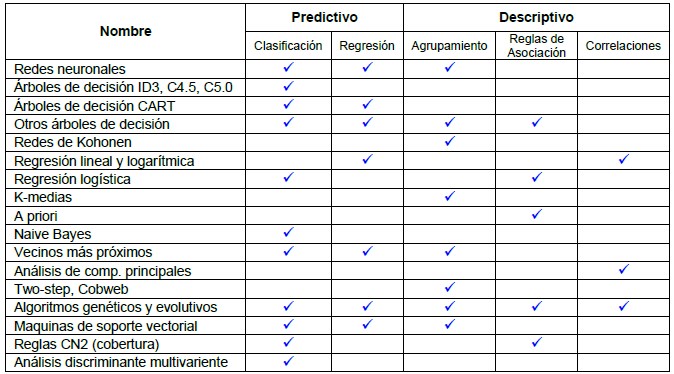


Fuente: Adaptado de Ruiz, C. A., Basualdo, M. S., & Pighin, S. (s. f.). Informática

Aplicada a la Ingeniería de Procesos I (Orientación I). *Data Mining*, 45.

A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas. Al ser más precisos y elaborados, se gana en capacidad predictiva, pero se pierde en rendimiento.

Tabla 6. Relación existente entre las tareas y técnicas de minería de datos.



Fuente: Hernández et al. (2004)

En la Tabla 6 se observa que todas las tareas pueden ser resueltas por más de una técnica, y que algunas técnicas pueden resolver más de una tarea, como es el caso de las redes neuronales, los árboles de decisión, las regresiones (lineal, logarítmica y logística) y los algoritmos genéticos, entre otros.

### 2.5.4 TIPOS DE MODELOS PARA ANÁLISIS PREDICTIVOS

El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos.

En el ámbito de los negocios los modelos predictivos extraen patrones de los datos históricos y transaccionales para identificar riesgos y oportunidades. Los modelos predictivos identifican relaciones entre diferentes factores que permiten valorar riesgos o probabilidades asociadas sobre la base de un conjunto de condiciones, guiando así al decisor durante las operaciones de la organización.

El efecto funcional que pretenden estas iniciativas técnicas es que el análisis predictivo provea una puntuación (probabilidad) para cada sujeto (cliente, empleado, paciente, producto, vehículo, componente, máquina y otra unidad en la organización) con el objeto de determinar, informar o influir procesos en la organización en el que participen un gran número de sujetos, tal y como ocurre en marketing, evaluación de riesgo de crédito, detección de fraudes, fabricación, salud y operaciones gubernamentales como el orden público.

Cuando se habla de análisis predictivo generalmente se quiere hablar de "modelos predictivos", datos de puntuaciones sobre la base de modelos predictivos y previsiones. No obstante se está generalizando el uso del término para relacionarlo con disciplinas analíticas y está muy extendido su uso para la segmentación entre usuarios de negocio y decisores.

2.5.4.1 MODELOS PREDICTIVOS

Los modelos predictivos son modelos de la relación entre el rendimiento específico de un sujeto en una muestra y uno o más atributos o características del mismo sujeto. El objetivo del modelo es evaluar la probabilidad de que un sujeto similar tenga el mismo rendimiento en una muestra diferente. Esta categoría engloba modelos en muchas áreas como el marketing, donde se buscan patrones de datos ocultos que respondan preguntas sobre el comportamiento de los clientes o modelos de detección de fraude.

Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un cliente o transacción en particular, de forma que aporte conocimiento a la hora de tomar una decisión. Gracias a los avances de ingeniería en el análisis de grandes volúmenes de datos estos modelos son capaces de simular el comportamiento humano frente a estímulos o situaciones específicas tal y como lo manifiesta Nyce, Charles (2007).

2.5.4.2 MODELOS DESCRIPTIVOS

En el modelo descriptivo se identifican patrones que describen los datos mediante tareas, ej. Agrupamiento (*clustering*) y reglas de asociación (Vazirgiannis et al., 2003). Hernández et al. (2004) destacan que mediante este modelo se identifican patrones que explican o resumen el conjunto de datos, siendo estos útiles para explorar las propiedades de los datos examinados.

Los modelos descriptivos siguen un tipo de aprendizaje no supervisado, que consiste en adquirir conocimiento desde los datos disponibles, sin requerir influencia externa que indique un comportamiento deseado al sistema (Sumathi y Sivanandam, 2006). Los modelos descriptivos cuantifican las relaciones entre los datos de manera que es utilizada a menudo para clasificar clientes o contactos en grupos. A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular, los modelos descriptivos identifican muy diferentes relaciones entre los clientes y los productos.

Los modelos descriptivos no clasifican u ordenan a los clientes por su probabilidad de realizar una acción particular de la misma forma en la que lo hacen los modelos predictivos. Sin embargo, los modelos descriptivos pueden ser utilizados por ejemplo para asignar categorías a los clientes según su preferencia en productos o su franja de edad. Eckerson, Wayne (2007) dice que las aplicaciones de los modelos descriptivos pueden ser utilizadas para desarrollar nuevos modelos adicionales que pueden imitar un gran volumen de agentes individuales y hacer predicciones.

2.5.4.3 MODELOS DE DECISIÓN

Los modelos de decisión están basados en la técnica de Dinámica de sistemas de Jay Forrester y describen la relación entre todos los elementos de una decisión; es decir los datos conocidos incluyendo los resultados de los modelos predictivos, la decisión y el pronóstico de los resultados de una decisión con el objeto de predecir los resultados de una decisión involucrando una gran cantidad de variables. Estos modelos pueden ser utilizados en la optimización o maximización de determinados resultados al mismo tiempo que otros son minimizados. Los modelos de decisión son generalmente usados para el desarrollo de la decisión lógica o conjunto de reglas de negocio que deberían producir el resultado deseado para cada cliente o circunstancia. Coker, Frank (2014).

### 2.5.5 HERRAMIENTAS DE ANÁLISIS PREDICTIVO

Utilizar análisis predictivo y entender sus resultados requería de los usuarios habilidades muy avanzadas en el pasado. Con la popularización del análisis avanzado las herramientas más modernas han mejorado sustancialmente el acceso a todo tipo de usuarios analistas. Según el informe *The Forrester Wave™[[15]](#footnote-15): Big Data Predictive Analytics Solutions, Q2 2015*: ―Organizaciones de todas las industrias se están animando a darle valor al análisis predictivo. Con el crecimiento de su demanda, los proveedores de análisis predictivo están proveyendo herramientas que reducen la barrera y aumentan el atractivo a aquellos con menos habilidades estadísticas‖. Al mismo tiempo que el análisis predictivo es utilizado en los procesos de toma decisiones y es integrado en los procesos de las organizaciones, también se está operando un cambio en el mercado en relación al tipo de usuario que consume la información siendo cada vez más relevante el papel del usuario de negocio.

Los usuarios de negocio demandan herramientas que ellos mismos puedan utilizar con autonomía según manifiesta Halper, Fern (2011). Los proveedores están creando nuevas aplicaciones que eliminan la complejidad matemática, proviniendo interfaz gráfica fácil de usar e incorporan accesos rápidos para, por ejemplo, reconocer qué tipo de datos están disponibles o sugerir un análisis predictivo apropiado. Las plataformas de análisis predictivo buscan con empeño la sofisticación que les permita analizar y visualizar la información oculta en sus datos de forma que cualquier cargo en la organización pueda extraer conocimiento útil para la toma de decisiones. La visualización de los datos se hace habitualmente a través de tablas, cuadros, gráficos y clasificaciones que muestran visualmente las clasificaciones más probables resultadas de las predicciones.

Hay una cantidad creciente de herramientas disponibles en el mercado que permiten la ejecución de análisis predictivo y se pueden ordenar desde aquellas que necesitan muy poca sofisticación del usuario hasta aquellas que están diseñadas para expertos y estadistas. También informa MacLennan, Jamie (2012) que se pueden clasificar por el nivel de personalización y el rendimiento en grandes volúmenes de datos.

2.5.5.1 HERRAMIENTAS DE CÓDIGO ABIERTO

Algunas herramientas de código abierto son:

* GNU Octave
* KNIME
* Orange
* R/Weka
* scikit-learn

2.5.5.2 HERRAMIENTAS DE CÓDIGO PRIVATIVO

Algunas herramientas comerciales son:

* Angoss KnowledgeSTUDIO
* KXEN Modeler
* Mathematica
* MATLAB
* Minitab
* Neural Designer
* OpenText™ Big Data Analytics

2.5.5.3 ESPECIFICACIÓN DEL LENGUAJE DE PROGRAMACIÓN

Para el desarrollo de la investigación se usa Python como lenguaje de programación.

Python es un lenguaje de programación interpretado dinámico y es multiplataforma.

La sintaxis hace que favorezca un código legible. Se trata de un lenguaje de programación multiparadigma, ya que soporta la programación orientación a objetos

*(POO)*. Es administrado por la Python[[16]](#footnote-16) Software Foundation.

Imagen 26. Logo Python



Fuente: Adaptado de IDEs para Python en Linux. (2015, noviembre 4). Recuperado 18 de abril de 2018, a partir de https://www.linuxadictos.com/las-tres-mejores-ides-python-de-

codigo-abierto.html

Gracias a que el intérprete de Python incluye un modo interactivo en el cual se escriben las instrucciones en una especie de *intérprete de comandos* las expresiones pueden ser introducidas una a una generando que el resultado se vea de forma inmediata, lo que da la posibilidad de probar fragmentos de código antes de integrarlo como parte de un programa final. Esto ha sido útil para el desarrollo del prototipo.

2.5.5.4 ESPECIFICACIÓN DE LAS LIBRERÍAS MACHINE LEARNING

SCIKITS.LEARN

Para la investigación se usa las librerías de *Scikit-learn* que son específicas y de gran importancia para la codificación y ejecución de los modelos y herramientas de Machine Learning. La definición general que dan los autores Fabian Pedregosa y Gaël Varoquaux de esta librería *Scikit-learn* es en resumen una biblioteca de aprendizaje de máquina de software libre para el lenguaje de programación Python.

Además agrupa dentro de su basta lista de algoritmos y modelos los del tipo *clasificación*, *regresión*, *agrupación*, *máquinas de vectores de soporte (SVN)*, *k-medias* entre otros. Está diseñado para interoperar con las bibliotecas numéricas y científicas de

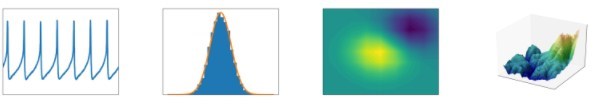
*Python[[17]](#footnote-17), NumPy* y *SciPy*.

* Matplotlib

Matplotlib es una biblioteca para la generación de gráficos a partir de datos contenidos en *listas* o *Arrays* en el lenguaje de programación *Python* y su extensión matemática *NumPy*.

Esta librería proporciona una API, pylab, diseñada para recordar a la de MATLAB.

Imagen 27. Opciones de gráficos Matplot



Fuente: Adaptado de Matplotlib: Python plotting — Matplotlib 2.2.2 documentation. (s. f.). Recuperado 19 de abril de 2018, a partir de <https://matplotlib.org/>

* Numpy

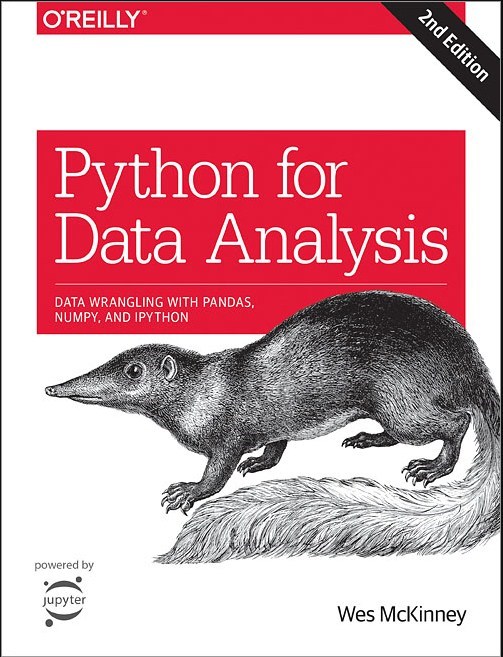
NumPy es una extensión de Python, que le agrega mayor soporte a la utilización de vectores y matrices, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar matrices. En 2005, Travis Oliphant creó NumPy[[18]](#footnote-18) incorporando características de *Numarray* en NumPy con algunas modificaciones. NumPy es open

source.

Uno de los autores más influyentes en la utilización de esta librería es Wes McKinney (s.

f.) quien en su libro *Python for Data Analysis, 2nd Edition Data Wrangling with Pandas, NumPy, and IPython* describe todo un completo instructivo para la manipulación, proceso y cruce de datos usando las librerías descritas en esta sección.

Imagen 28. Libro Python for Data Analysis



Fuente: Adaptado de McKinney, W. (s. f.). Python for Data Analysis. Recuperado el 31 de marzo del 2018 del sitio web <http://shop.oreilly.com/product/0636920050896.do>

 Pandas

Pandas es una biblioteca de software escrita como extensión de NumPy para manipulación y análisis de datos para el lenguaje de programación Python.

En particular, ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales. Es un software libre distribuido bajo la licencia BSD [[19]](#footnote-19) versión tres cláusulas.

2.5.5.5 ESPECIFICACIÓN DE HERRAMIENTAS

A continuación se describen las herramientas que se utilizan de Minería de Datos en la investigación:

* Bottle

Como entorno visual para diseñar el prototipo se usa este framework que es ideal para el desarrollo de aplicaciones web, es muy simple y fácil de usar ya que proporciona un conjunto de herramientas para los desarrolladores como plantillas, accesos rápidos de comandos. Bottle se puede ejecutar en Python 3.

Imagen 29. Framework Bottle



Fuente: Adaptado de Bottle: Python Web Framework — Bottle 0.13-dev documentation.

(s. f.). Recuperado 13 de julio de 2018, de https://bottlepy.org/docs/dev/

* Jupyter

La aplicación Jupyter Notebook es una aplicación cliente-servidor de código abierto que permite editar y ejecutar documentos que contengan códigos, ecuaciones, visualizaciones y textos a través de un navegador web. La aplicación Jupyter Notebook se puede ejecutar en un escritorio local que no requiere acceso a Internet. Los usos incluyen: limpieza y transformación de datos, simulación numérica, modelado estadístico, visualización de datos, aprendizaje automático.

Imagen

30

. Logo de Jupyter



Fuente: Adaptado de Project Jupyter. (s. f.). Recuperado 20 de abril de 2018, del sitio

web Jupyter: http://www.jupyter.org

* Anaconda

Anaconda es una distribución de código abierto para hacer ciencia de datos y aprendizaje automático, es por ese motivo que es el referente utilizado en la investigación. Incluye más de 250 paquetes populares de ciencia de datos, el paquete *conda[[20]](#footnote-20)* y el administrador de entorno virtual para Windows, Linux y MacOS. Conda hace que sea fácil y rápido instalar, ejecutar y actualizar entornos complejos de ciencia de datos y aprendizaje de máquinas como Scikit-learn, TensorFlow y SciPy.

Esta aplicación es la base de millones de proyectos de ciencias de datos, así como del aprendizaje automático de Amazon Web Services y Anaconda para Microsoft en Azure y Windows.

Imagen

31

. Distribución Anaconda



Fuente: Adaptado de Anaconda. (s. f.). What is Anaconda? Recuperado 20 de abril de 2018, del sitio web Anaconda: <https://www.anaconda.com/what-is-anaconda/>

* PostgreSQL

PostgreSQL es un sistema de gestión de bases de datos relacional orientado a objetos y libre. PostgreSQL provee soporte para

* Números de precisión arbitraria
* Texto de largo ilimitado
* Figuras geométricas (con una variedad de funciones asociadas)
* Arrays

Imagen 32. Base de datos PostgreSQL



Fuente: adaptado de Luz, S. D. (2016, febrero 13). PostgreSQL Recuperado 27 de abril de 2018, del sitio web redeszone: [https://www.redeszone.net/2016/02/13/postgresqlhttps://www.redeszone.net/2016/02/13/postgresql-lanza-una-actualizacion-de-seguridad-critica-para-todas-sus-versiones-con-soporte/lanza-una-actualizacion-de-seguridad-critica-para-todas-sus-versiones-con-soporte/](https://www.redeszone.net/2016/02/13/postgresql-lanza-una-actualizacion-de-seguridad-critica-para-todas-sus-versiones-con-soporte/)

* Talend Studio

Talend es una herramienta muy poderosa de código abierto para la integración de datos permite el armado y la configuración de procesos de extracción y carga de archivos.

Imagen 33. Talend Studio



Fuente: Adaptado de Talend, líder de integración de datos en la nube (ETL moderno). (s. f.). Recuperado 23 de junio de 2018, del sitio web Talend: <https://es.talend.com/>

El análisis del reporte del cuadrado mágico de Gartner[[21]](#footnote-21) para herramientas de integración lo ubicó como líder por segunda vez consecutiva como una de las mejores herramientas de software libre para la integración de datos.

# 3. METODOLOGÍA DE LA INVESTIGACIÓN

## 3.1 METODOLOGÍAS DE MINERÍA DE DATOS

Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Estas metodologías ayudan a entender el proceso de descubrimiento de conocimiento para proveer una guía de planificación y ejecución en los proyectos.

Algunos modelos conocidos como metodologías son en realidad un modelo de proceso definiendo un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo.

Por esta razón, diversas empresas y consultorías en el mundo han desarrollado metodologías de trabajo para guiar al usuario para una aplicación exitosa de las técnicas de minería de datos a través de una sucesión de pasos (Sumathi y Sivanandam, 2006).

La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer y la metodología especifica cómo hacerlo.

Dentro de las metodologías más adecuadas para la planificación de proyectos se encuentran las siguientes.

### 3.1.1 METODOLOGÍA SEMMA

SEMMA (Sampling, Exploration, Modification, Modeled, Assessment) desarrollado por la empresa Statistical Analysis Systems Inc. (SAS Institute), quien la define como:

―el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos‖.

3.1.1.1 FASES DE LA METODOLOGÍA SEMMA

Las fases de SEMMA[[22]](#footnote-22) y las tareas relacionadas son las siguientes[[23]](#footnote-23):

* Muestra:

El proceso comienza con el muestreo de datos, por ejemplo, seleccionando el conjunto de datos para modelar.

El conjunto de datos debe ser lo suficientemente grande como para contener suficiente información para recuperar, pero lo suficientemente pequeño como para ser utilizado de manera eficiente. Esta fase también se ocupa de la partición de datos.

* Explora:

Esta fase cubre la comprensión de los datos al descubrir relaciones anticipadas e imprevistas entre las variables, y también anomalías, con la ayuda de la visualización de datos.

* Modificar

La fase de modificación contiene métodos para seleccionar, crear y transformar variables en preparación para el modelado de datos.

* Modelo

En la fase de Modelo, la atención se centra en la aplicación de diversas técnicas de modelado (extracción de datos) en las variables preparadas con el fin de crear modelos que posiblemente proporcionen el resultado deseado.

* Evaluar:

La última fase es Evaluar. La evaluación de los resultados del modelado muestra la fiabilidad y la utilidad de los modelos creados.

### 3.1.2 METODOLOGÍA KDD

La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como Knowledge Discovery in Databases (KDD), que descubre conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información informa Han, J.; Kamber M. (2001).

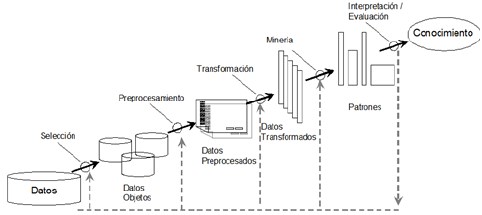
No es un proceso automático, es un proceso repetitivo que explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos. Klösgen W. y Zytkow J. M. (2002) definen al KDD como ―el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos‖.

Con base en la definición de Fayyad et al. (1996), Vazirgiannis et al. (2003) consideran que en el KDD los datos vienen a ser el conjunto de casos (registros) que pertenecen a la base de datos y un patrón es un subconjunto de datos que comparten propiedades

similares.

Imagen 34. Etapas de KDD

Fuente: Chapman et al., 2000



3.1.2.1 ETAPAS DEL PROCESO KDD

Se definen las etapas del proceso del descubrimiento de información:

* Selección de datos: En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

* Pre-procesamiento: Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores.

* En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

* Transformación: Consiste en el tratamiento preliminar de los datos,

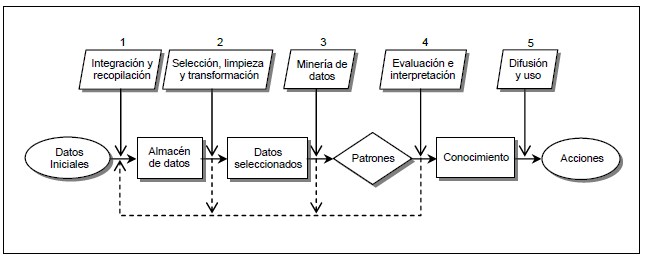
transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

* Minería de datos: Es la fase de modelamiento propiamente en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u ―ocultos‖ en los datos.

* Interpretación y evaluación: Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Asimismo, Fayyad et al. (1996) consideran que el KDD es un proceso iterativo e interactivo dividido en una secuencia de pasos, de los cuales la minería de datos es considerada como uno de los pasos más importantes en todo el proceso, definiéndolo como ―el corazón del proceso KDD mismo‖.

Imagen 35. Proceso de descubrimiento de conocimiento en bases de datos (KDD).



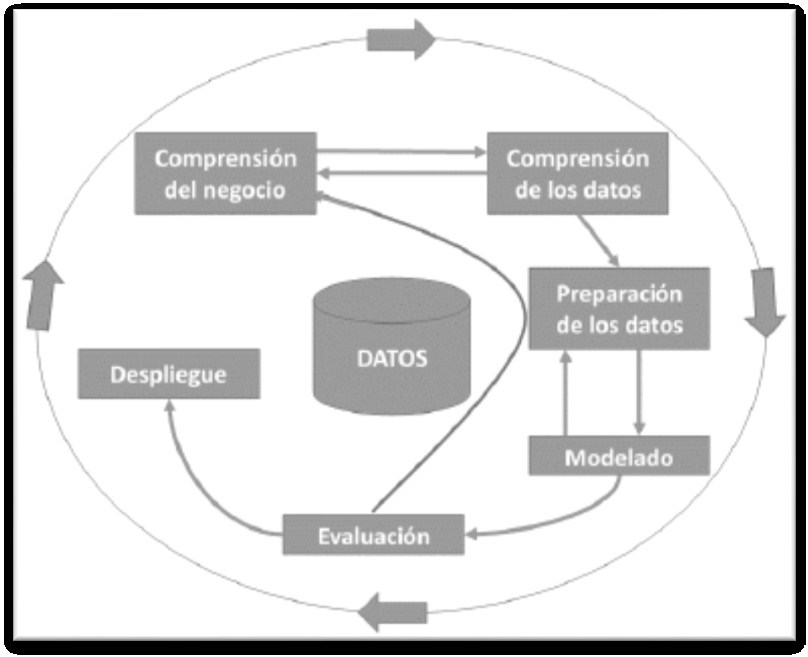
Fuente: Adaptado de Vazirgiannis et al. (2003) y Hernández et al. (2004)

### 3.1.3 METODOLOGÍA CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining), (Chapman et al., 2000) desarrollada en 1996 por el consorcio de empresas europeas, DaimlerChrysler, SPSS

(Statistical Product and Service Solutions) y NCR Systems Engineering Copenhagen. Esta metodología se encuentra dentro de las más utilizadas para la elaboración de proyectos de minería de datos, está basado en actividades ordenadas en seis fases que recorren todo el proceso de minería de datos, desde la definición de los objetivos del negocio que se pretende obtener, hasta la vigilancia y el mantenimiento del modelo que se proponga e implemente.

Imagen 36. Metodología CRISP-DM

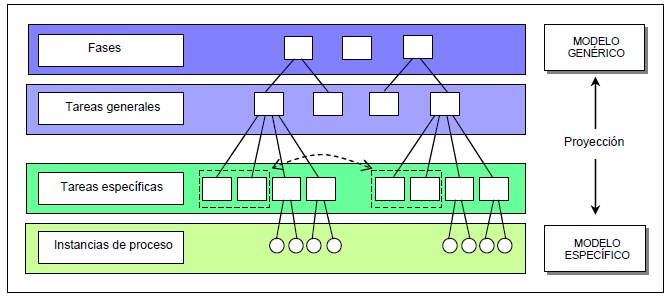


Fuente: Fuente: Chapman et al., 2000.

Chapman plantea que esta metodología consta de cuatro niveles, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. En el nivel general, el proceso está organizado en seis fases: análisis del problema, análisis de datos, preparación de datos, modelado, evaluación e implementación.

A su vez, cada fase consta de tareas generales de un segundo nivel. Estas tareas generales se proyectan a tareas específicas, donde se describen las acciones que se deben desarrollar para cada situación específica.

Imagen 37. Esquema de los cuatro niveles de abstracción de la metodología CRISPDM[[24]](#footnote-24).



Fuente: Fuente: Chapman et al., 2000.

3.1.3.1 FASES DE LA METODOLOGÍA CRISP-DM

Se definen las fases a continuación:

* Fase de comprensión del negocio: Se centra en la comprensión de los objetivos del proyecto de minería de datos desde un punto de vista de negocios. Esta fase es equivalente, por lo tanto, a una fase de análisis de requerimientos de un proyecto de desarrollo de software, y es importante porque que el cliente puede no tener

claro qué es lo que quiere.

Las tareas a realizar en esta fase incluyen determinar los objetivos de negocios, evaluar la situación del proyecto en términos de recursos, restricciones y suposiciones, determinar objetivos de minería de datos que traduzcan a criterios técnicos los objetivos de negocios y, finalmente, producir el plan del proyecto.

* Fase de comprensión de los datos: Comprende la recolección inicial de datos, identificando la calidad de estos y estableciendo las relaciones más evidentes entre ellos. Incluye la tarea de recolección de datos iniciales, en la cual deben inscribirse los datos en términos de número de registros, número de campos por registro y significado de cada campo. Incluye también la tarea de descripción de los datos en términos de tipo, distribución, tablas de frecuencia y estadísticas. Tareas adicionales de esta fase son la exploración de los datos mediante gráficos y tablas, y la verificación de la calidad de los mismos.

* Fase de preparación de los datos: En esta fase debe construirse una base de datos, la cual debe contener todas las características consideradas candidatas para estimar el valor de una variable que se espera predecir. Esta fase incluye la tarea de selección de los datos a los que se va aplicar la técnica del modelo, la tarea de limpieza de los mismos para alcanzar el nivel de calidad requerido por las técnicas de minería de datos que sean seleccionadas, la tarea de construir datos adicionales, la tarea de integrar diferentes bases de datos, y la tarea de formatear los datos.

La fase de preparación de los datos debe entregar datos que estén en un formato adecuado para la técnica del modelo que se empleará en la siguiente fase. Por esto, la fase de modelado puede requerir regresar una o más veces a la fase de preparación de los datos.

* Fase de modelado: Esta es la fase medular de un proyecto de minería de datos y consiste en descubrir una relación entre un conjunto de variables y una variable que se espera predecir. Contempla la selección de una técnica de modelado, entre las cuales pueden mencionarse las redes de KOHONEN o modelos K-MEAN para agrupamiento, árboles de decisión C5 o C&R para segmentación, redes neuronales o regresión logística para predicción, inducción de reglas generalizadas para descubrimiento de patrones y análisis de factores para reducir la complejidad de los datos, entre otras.

Entre sus tareas se encuentran la selección de la técnica del modelo, la generación del diseño de las pruebas del modelo, la construcción del mismo y, finalmente, la evaluación técnica a la que debe someterse el modelo a través de criterios

estadísticos.

* Fase de evaluación de negocios: Se evalúa el modelo de forma técnica en relación a factores tales como su precisión y generalidad. En esta fase, en cambio, debe evaluarse el nivel de satisfacción de los objetivos de negocios perseguidos por el proyecto de minería de datos.

Incluye la tarea de evaluar los resultados, la tarea de revisar el proceso de minería de datos y, finalmente, la tarea de determinar los próximos pasos a seguir (momento en el que debe decidirse si debe darse por terminado el proyecto de minería de datos y entrar en la fase de despliegue, si deben iniciarse iteraciones adicionales, o si debe iniciarse un nuevo proyecto de minería de datos).

* Fase de despliegue del modelo: En esta fase deberá definirse una estrategia para implementar los resultados de la minería de datos. Incluye las tareas de planificar el despliegue del modelo, de planificar el monitoreo y el mantenimiento de los modelos, de generar el reporte final del proyecto, y de revisar el proyecto en relación a evaluar lo que ocurrió correctamente y lo que necesita ser mejorado.

## 3.2 COMPARACIÓN DE METODOLOGÍAS

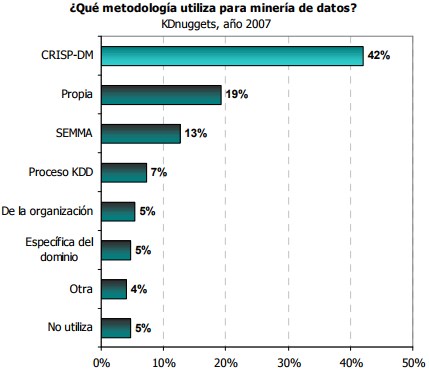
En el campo en expansión de la minería de datos, se ha solicitado una metodología estándar o simplemente una lista de mejores prácticas para el proceso diversificado e iterativo de minería de datos que los usuarios pueden aplicar a sus proyectos de minería de datos, independientemente de la industria. Mientras que el proceso estándar de Cross

Industry para Minería de Datos o CRISP-DM, fundado por la iniciativa del Programa Estratégico Europeo de Investigación en Tecnología de la Información, tenía como objetivo crear una metodología neutral, SAS también ofreció un patrón a seguir en sus herramientas de minería de datos.

Algunos modelos como dicen los siguientes autores Moine, I. J. M., Haedo, D. A. S., & Gordillo, D. S profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o

SEMMA).

Imagen 38. Encuesta realizada por la KDnuggets en el año 2007



Fuente: Adaptado de KDnuggets (s. f.). Encuesta de Metodologías. Recuperado el 15 de

febrero de 2018 del sitio web <https://www.kdnuggets.com/>

Tabla 7. Analogía entre las metodologías KDD y CRISP-DM



Fuente: Elaboración propia

## 3.3 RECURSOS HUMANOS

Las empresas deben lograr la máxima eficacia y eficiencia en sus empleados determinando necesidades de competencia y formación para el personal evaluando constantemente las acciones y los objetivos pactados por las gerencias.

Asegurar que el personal sea y esté consciente de la importancia de sus actividades y como deben contribuir a alcanzar el objetivo de satisfacción al cliente.

Algunos de los perfiles son:

* Educación: es necesario que se disponga de estudios mínimos como primaria, secundaria, universidad o no es imprescindible.

* Formación específica para dicho puesto, la cual es necesaria previo al inicio de la actividad, especialización en determinadas herramientas informáticas, certificados de manipulación de alimentos, manejo de armamento, manejo de maquinaria pesada.

* Habilidades de contacto y fluidez verbal con el usuario final.

* Experiencia mínima en el manejo de ciertas circunstancias que impliquen una rápida toma de decisión.

Las empresas deben identificar, proporcionar y mantener la infraestructura necesaria para lograr la conformidad del servicio y para el empleado.

* Espacio y medios de trabajo

* Equipos apropiados incluyendo hardware y software

* Instalaciones asociadas

* Servicios de apoyo como transporte, alimentación y comunicación.

# 4. DISEÑO DE LA SOLUCIÓN

La toma de decisiones implementada con inteligencia de negocios, a través de herramientas de minería de datos, contribuirá de gran manera a una mejor planeación en el área administrativa, de recursos humanos y gerencia, para evitar el abandono laboral y apoyar en todo momento al empleado. Con el objetivo de buscar los factores de deserción de los empleados, en este trabajo se aplican técnicas de minería de datos apoyándose en la metodología CRISP-DM. En esta fase se comprende la recolección y comprensión de los datos, validar la calidad de los mismos, depurar los datos y establecer una relación entre los campos y atributos provenientes de diferentes fuentes de datos que se utilizarán más adelante.

La investigación consiste en implementar un proceso predictivo para el abandono laboral; aplicando minería de datos basado en técnicas de clustering*.* Como principal fuente de datos para llevar a cabo esta investigación en la empresa existe un registro de más de 11 años y con más de 5000 registros. Al iniciar cada año, depende de la cantidad de nuevos proyectos contratados pueden ingresan en promedio 180 empleados de lo contrario pueden egresan un promedio de 230 empleados con relación al año pasado. Las variables a estudiar, por empleado, son: que género tiene, la edad, sueldo, en que empresa está y en cual estará para un empleado nuevo, número de hijos, con cuántos hijos en su residencia convive, cantidad de años de experiencia laboral posee, si es casado o soltero, si es fumador, si habla otro Idioma, conocimiento de programas de computador, posee auto, sabe manejar automóviles.

Como principal herramienta, para este proyecto se utiliza Anaconda Distribution 5.0.1 ya que es la plataforma más popular de software libre para el manejo científico de datos usando Python como principal lenguaje de programación y Framework Bottle como entorno web del prototipo.

## 4.1 ANÁLISIS DE DATOS

En esta fase se integraron, transformaron y normalizaron los sets de datos de las 5 empresas-sedes *(objetivos)* para la ejecución en el prototipo utilizando las técnicas de minería de datos descritas en los párrafos anteriores.

La fase de análisis de datos comprende la recolección de los datos de los empleados de la empresa *Argentina de Seguridad* a la cual estamos haciendo el caso de estudio respectivo y que fueron obtenidos desde diferentes sistemas de información de la empresa. Las variables a estudiar para cada empleado son: *Sexo, Edad, Sueldo, Empresa, Hijos,*

*Hijos\_Residencia, Experiencia\_Laboral, Casado, Fumador, Otro\_Idioma, Maneja\_Pc,*

*Posee\_Auto, Sabe\_Manejar, Tiempo Días.*

En ―*The 7 Hidden Reasons Employees Leave*‖ el especialista Leigh Branham declara que existen problemas internos en las organizaciones que obligan a los empleados a buscar más movilidad laboral y generando altos grados de deserción en la empresa, estas son algunas de las causas:

Tabla 8. Factores de deserción con valores de la empresa.

|  |
| --- |
| El empleado se crea expectativas del trabajo. |
| La persona no coincide con el puesto |
| Poco seguimiento al empleado |
| Pocas oportunidades de crecimiento |
| Desconfianza en altos mandos |

Fuente: Elaboración propia

Inicialmente se tiene una base de datos de empleados, de un periodo de más de 10 años la cual ayudará a encontrar los principales indicadores así como aplicar correctamente los algoritmos de minería de datos y poder validar que las variables a utilizar sean las correctas. Esta información será relevante para poder lograr los objetivos de la tesis.

## 4.2 PREPARACIÓN DE LOS DATOS

En esta etapa intervienen una serie de pasos y componentes fundamentales que acompañan el proceso previo para implementar correctamente las técnicas de minería de datos. El primer paso comprende la comprensión del dominio del caso de estudio junto con el detalle correcto de los objetivos puesto que es fundamental tener muy claros los límites de la investigación y de esa forma no perder el rumbo.

Para el segundo paso es clave para la creación del *set de datos* la correcta selección e integración de los datos provenientes de diferentes fuentes, los datos que se necesitan están esparcidos por toda la empresa y son disparejos en varios sectores, estos van desde colecciones de documentos, correos electrónicos, fotografías, clips de vídeo, bases de datos de procesos, bases de datos de transacciones de clientes, registros web, o los *logs* de las páginas web, etc. Es importante homogeneizar todos los formatos de los datos para que la información sea más fácil de procesar y analizar.

Dando continuidad al tercer paso limpieza y procesamiento de datos es importante eliminar todo ruido y datos aislados como inconsistencias y duplicidad de contenido, los ―datos no depurados‖ confunden el proceso de minería y producen a resultados inválidos o poco confiables. Los pasos implementados hasta este momento tienen el objetivo de mejorar la calidad de los datos y los resultados de la utilización de técnicas y herramientas de minería de datos en el proceso de esta investigación.

### 4.2.1 PROCESO KDD KNOWLEDGE DISCOVERY IN DATABASES

*“La tecnología KDD está basada en un bien definido proceso KDD de múltiples pasos, para el descubrimiento de conocimiento en grandes colecciones de datos. El proceso KDD es iterativo por naturaleza, y depende de la interacción para la toma de decisiones, de manera dinámica”.* [[25]](#footnote-25)

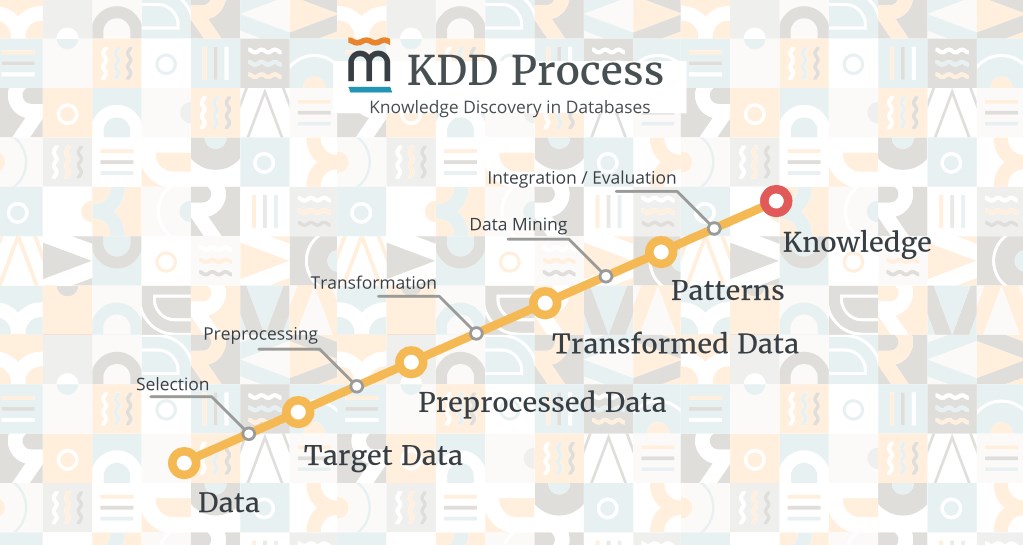
Con esta definición la intención de la investigación es elaborar un sistema *(prototipo)* de descubrimiento de conocimiento y toma de decisiones con una interfaz simple y fácil de utilizar. Actualmente se puede encontrar un gran número de definiciones acerca de este proceso de descubrimiento de conocimiento en las bases de datos.

Una de las definiciones más completas, es la siguiente: *“El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos*)‖ Usama Fayyad y Evangelos Simoudis. (1997)

Este marco de trabajo compuesto de varias etapas, actualmente está siendo utilizado por más empresas debido al crecimiento de la información almacenada en las bases de datos y la inversión en hardware para la capacidad de procesar los datos.

El siguiente gráfico muestra el paso a paso de las etapas en el camino hacia el descubrimiento del conocimiento en los datos de la empresa.

Imagen 39. Proceso KDD



Fuente: Adaptado de Data, M. (s. f.). KDD: Knowledge Discovery in Databases.

Recuperado 24 de marzo de 2018, a partir de <http://mnrva.io/kdd-platform.html>

La primera etapa consiste en la selección de las fuentes de datos relevantes para el análisis de los mismos, pasando por el pre-proceso de limpieza y preparación de los datos generando una única estructura adecuada para la transformación de los mismos. Culminando estas etapas se implementan las herramientas y modelos de minería de datos para obtener los patrones ocultos y/o existentes en la información almacenada, dando continuidad a la interpretación, análisis y evaluación de los resultados.

En la *sección 3* se amplía todo el proceso referente a las metodologías para el descubrimiento de datos en la información que tiene la empresa.

## 4.3 CREACIÓN DE LA BASE DE DATOS

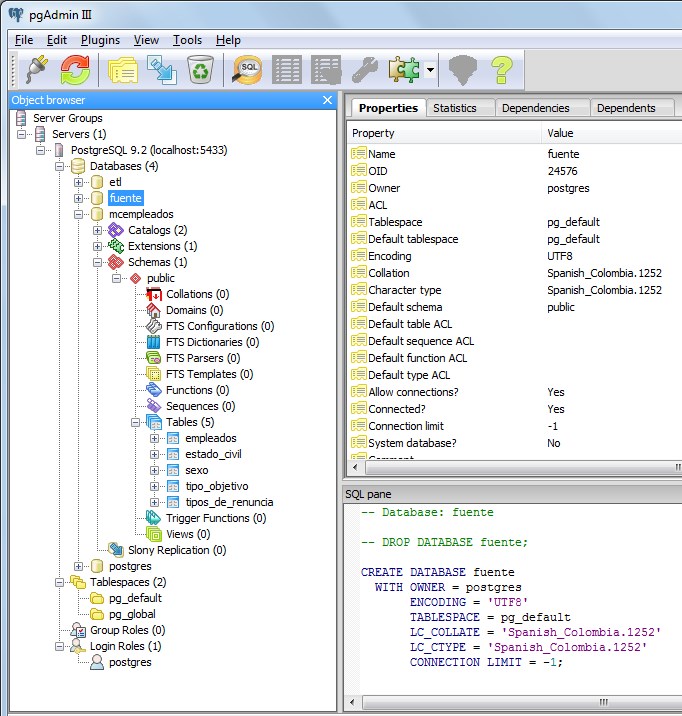
El siguiente paso que se usa de la metodología CRISP-DM es de preparación de los datos. En esta fase se hace necesario construir una base de datos, la cual va a contener todas las características consideradas candidatas para estimar el valor de una variable que se espera predecir.

Se incluyen tareas de selección de los datos a los que se va aplicar la técnica del modelo, la tarea de limpieza de los mismos para alcanzar el nivel de calidad, la tarea de construir datos adicionales y la tarea de dar formato a los datos.

Una vez que se obtiene la estructura final con la información, se procede con la transformación de los datos al gestor *PostgreSQL*. De esta manera se crea la base de datos para los modelos de minería de datos.

Ilustración 1. Tablas sexo, tipo\_objetivo, tipos\_de\_renuncia

Fuente: Elaboración propia



## 4.4 CONSTRUCCION DEL ETL

Para la construcción del proceso del ETL se usó la aplicación Talend Studio, el cual facilita el proceso de la integración de los datos de la empresa del caso de investigación. En las siguientes secciones se definen y explican los procesos realizados para dicho procedimiento.

### 4.4.1 PREPARAR LOS DATOS ETL

ETL por sus siglas significa extraer, transformar y cargar. Es uno de los procesos importantes dentro de las empresas para mover datos desde múltiples fuentes, cambiar el formato y depurar datos innecesarios para cargarlos en otra base de datos, data mart, o DataWarehouse, este proceso se usa para el análisis de metadatos o para apoyar un proceso de negocio. En la sección *2.4 clasificación de bases de datos* se extiende la definición de ETL.

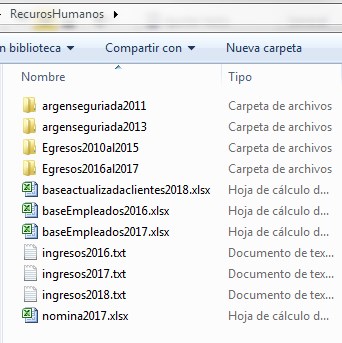
### 4.4.2 SELECCIÓN DE LA FUENTE DE DATOS

Se trabajó con datos de diferentes fuentes y repositorios, se necesitó pre procesar y preparar los datos antes del desarrollo de modelos predictivos. Se detectó que los datos carecían de ciertos valores y que estos eran erróneos y tenían diferentes formatos en su estructura.

* Al desarrollar la investigación se encontraron algunas limitaciones, no solo en el acceso a los datos, sino también en la interpretación de los mismos.
* Una primera limitación proviene del acceso a los datos y documentos, los mismos no fueron proporcionados en su totalidad por la empresa. La confidencialidad de dicha información fue una característica principal para el acceso a los datos.

La primera parte del proceso ETL consistió en extraer los datos desde los sistemas de origen, para la investigación y caso de estudio de la empresa, los datos provienen de carpetas de los empleados del area de recursos humanos, archivos planos en *txt*, archivos de *Excel* y la exportación del sistema de nómina del area contable.

Imagen 40. Fuente de datos

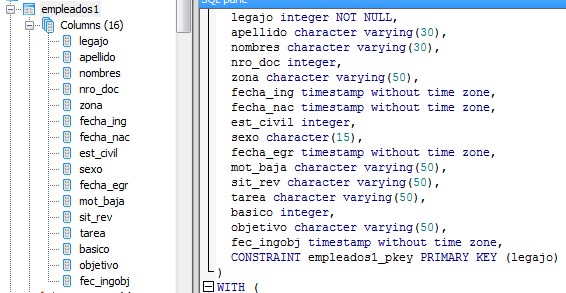


Fuente: Elaboración propia

Una vez que se identificaron los orígenes y fuentes de datos, se procede a la construcción de las tablas físicas del modelo y al desarrollo de los procesos de carga. Se generan todos los registros necesarios con un periodo de 10 años, que va desde el 01 de Enero de 2007 hasta el 31 de Diciembre de 2017.

Imagen 41. Diseño físico de la tabla

Fuente: Elaboración propia



Se aplican algunas de las siguientes transformaciones:

* Seleccionar y depurar ciertas columnas para su carga (por ejemplo, la columna con número de documento, número de legajo no se cargan).

* Traducir códigos de las columnas género y tipo de género (por ejemplo, si la fuente almacena una ―H‖ para Hombre y ―M‖ para Mujer o el destino tiene que guardar ―1″ para Hombre y ―2″ para Mujer).

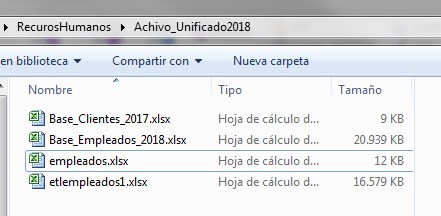
* Obtener nuevos valores calculados (por ejemplo, total\_años\_trabajaos =

cantidad\_dias / 365).

* Dividir una columna en varias (por ejemplo, columna ―Nombre: García, Miguel‖; pasar a dos columnas ―Nombre: Miguel‖ y ―Apellido: García‖).

Para trabajar con los datos de los objetivos *(empresas que adquieren el servicio de vigilancia)* obtenidos se debe considerar como primera instancia todos los archivos y fuentes de datos como se detalla en la siguiente imagen.

Imagen 42. Carpeta datos a unificar



Fuente: elaboración propia

Para lo cual se integró la información de las mismas en un archivo único conformado por los campos que se presentan en la siguiente Tabla 9.

Tabla 9. Campos disponibles para la preparación de datos

|  |  |  |
| --- | --- | --- |
| Campo | Descripción | Tipo |
| Sexo | Identifica el género de la persona | Entero |
| Edad | Identifica la edad de la persona | Entero |
| Sueldo | Registra el valor a contratar | Entero |
| Empresa | Identifica la sucursal para la persona | Entero |
| Hijos | Registra el número de hijos | Entero |
| Hijosresidencia | Registra el número de hijos que conviven | Entero |
| Experiencia | Registra el número de años de experiencia | Entero |
| Casado | Registra el estado civil de la persona | Entero |
| Fumador | Identifica si fuma la persona | Entero |
| Otroidioma | Identifica si la persona sabe otro idioma | Entero |
| Manejapc | Identifica si la persona maneja computador | Entero |
| Poseeauto | Identifica si la persona posee automóvil | Entero |
| Sabemanejar | Identifica si la persona sabe manejar automóvil | Entero |
| Tiempo días | Calcula el número de días para la persona | Entero |

Fuente: Elaboración propia

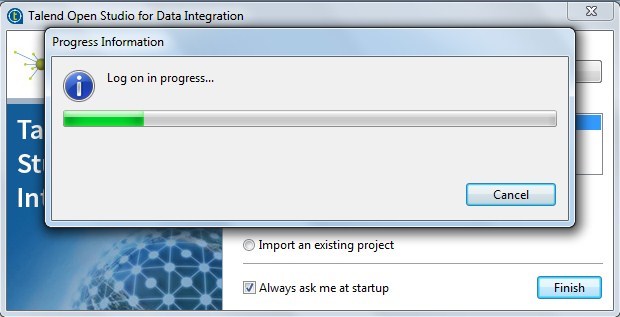
La fase de carga va a interactuar directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones que se definieron anteriormente, por ejemplo, valores únicos, integridad referencial, campos obligatorios, rangos de valores.

Estas restricciones contribuyen a que se garantice la calidad de los datos en el proceso

ETL y deben ser tenidos en cuenta.

Para la construcción del ETL se utiliza la herramienta Talend Studio.

Ilustración 2. Carga del proyecto

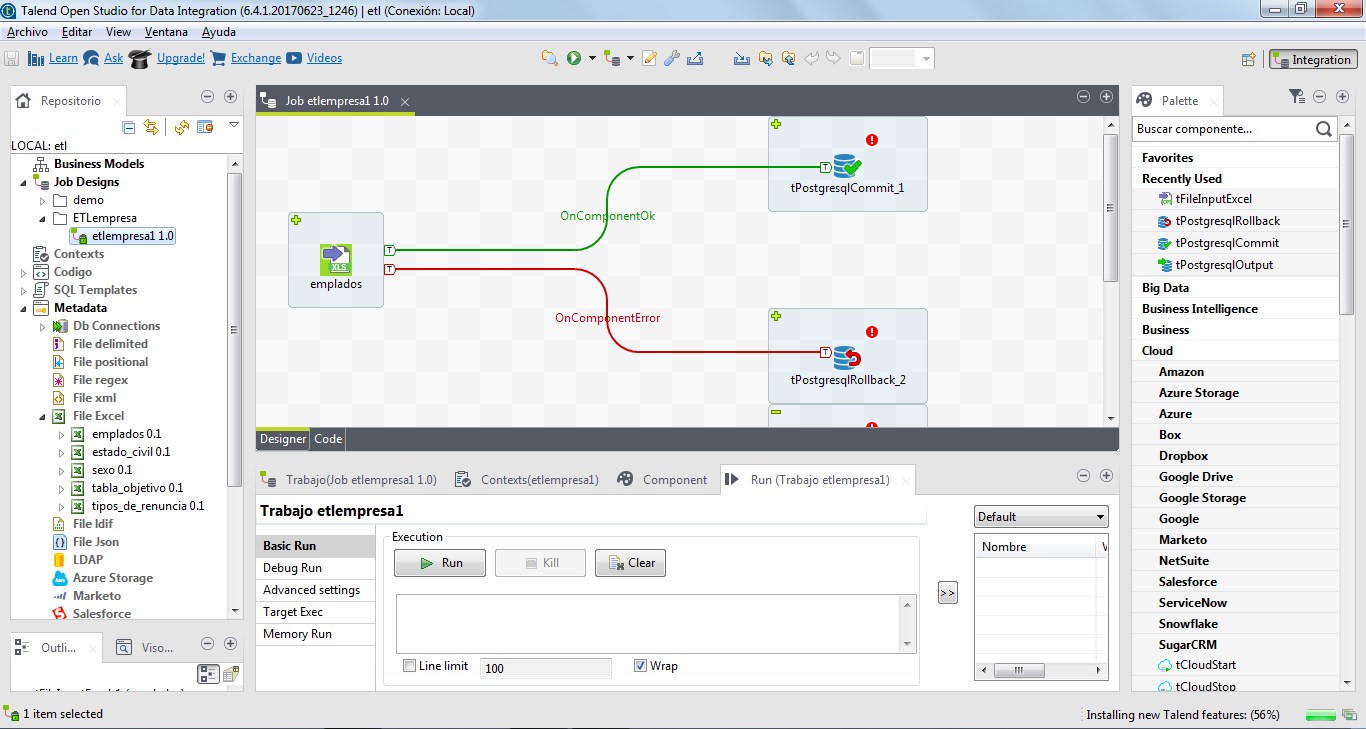


Fuente: Elaboración propia

El segundo paso es identificar las fuentes de datos ya *normalizadas* y *depuradas* para cargarlas en la herramienta, generando en la conexión a la base de datos el respectivo *commit* y *rollback*.

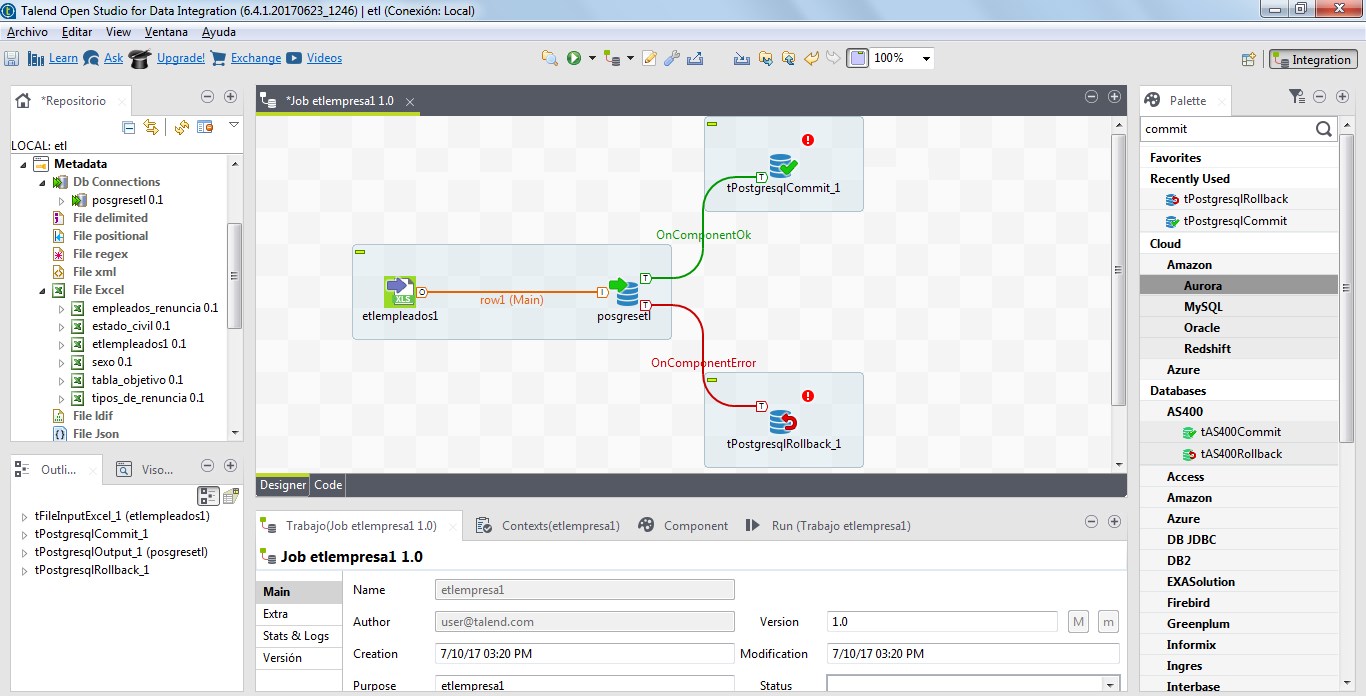
Ilustración 3. Configuración con las fuentes de datos

Fuente: Elaboración propia



Se realiza la conexión con la base de datos:

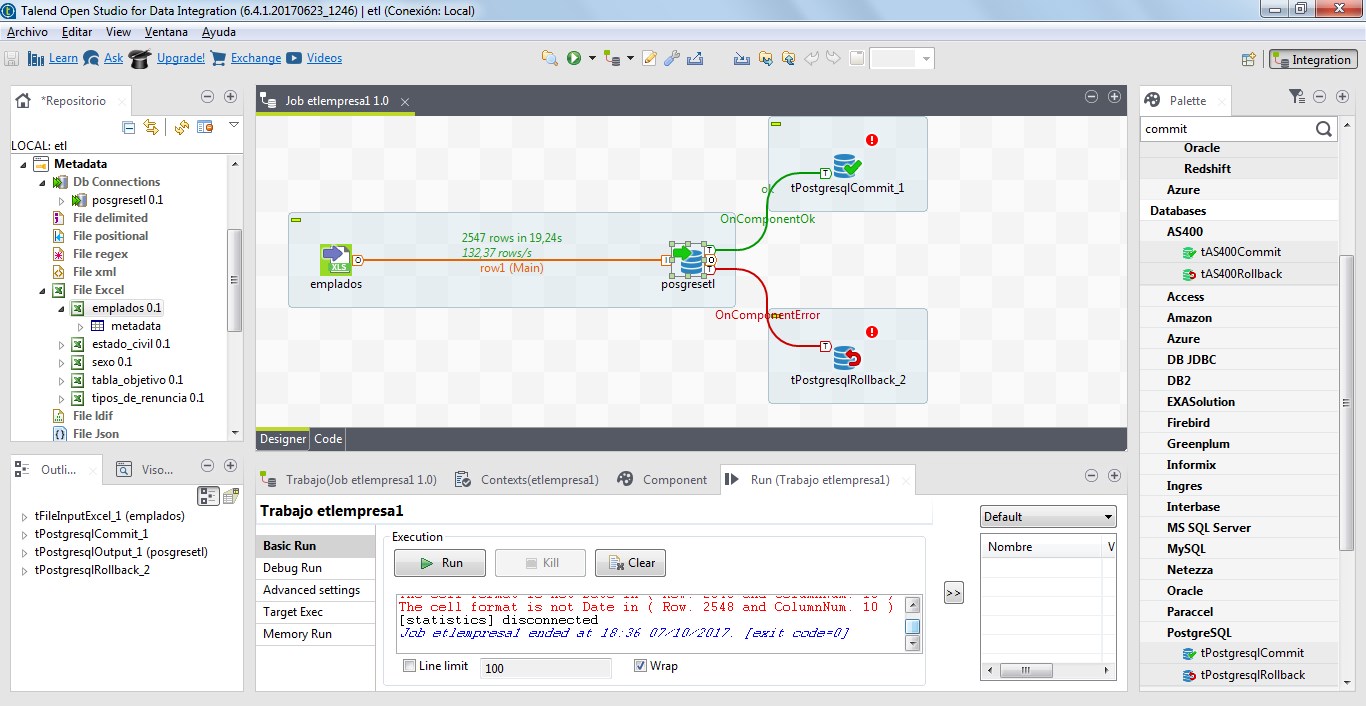
Ilustración 4. Conexión con las fuentes de datos



Fuente: Elaboración propia

Se realiza correctamente la conexión a las fuentes de datos.

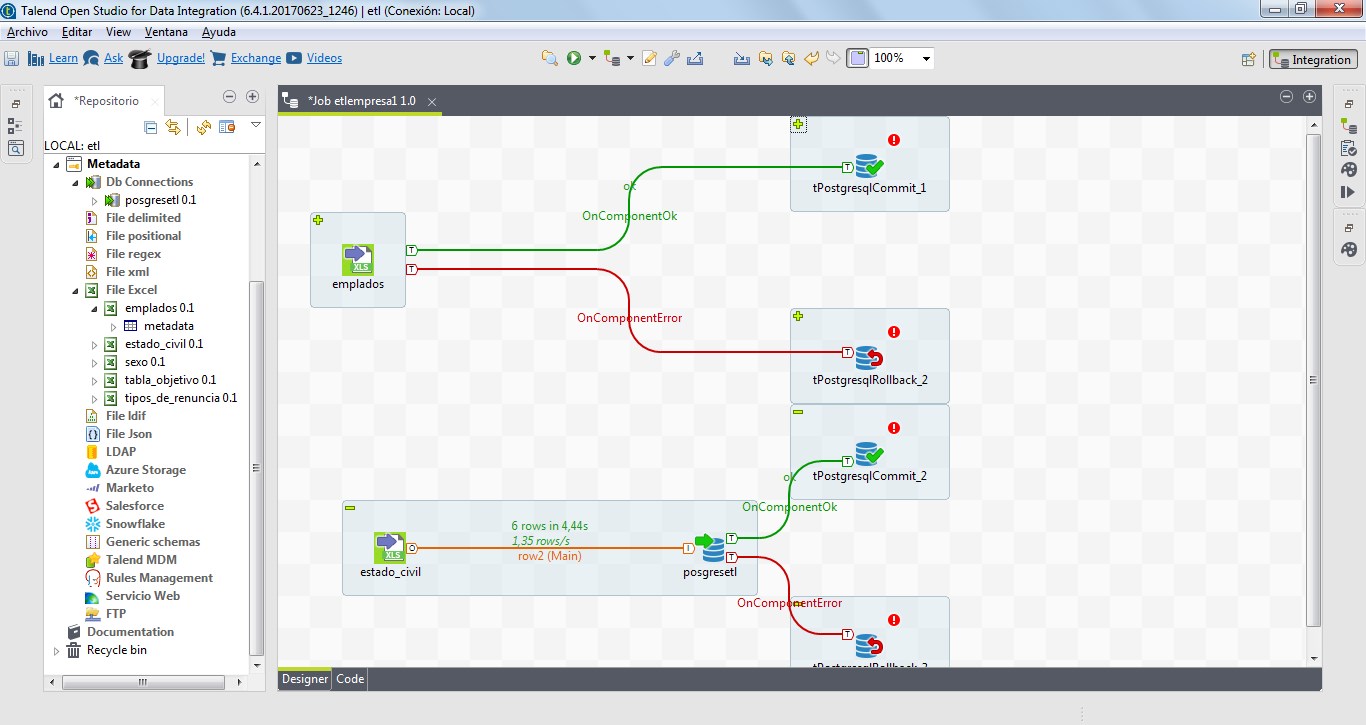
Ilustración 5. Prueba de conexión ―Exitoso‖ con las fuentes de datos



Fuente: Elaboración propia

Se realiza la conexión con las fuentes de datos.

Ilustración 6. Conexión con las fuentes de datos



Fuente: Elaboración propia

### 4.4.3 CONSTRUCCION DEL DATAWAREHOUSE

Con el resultado de los dos procesos anteriores se arma el archivo final con el resultado del consolidado de la base de datos. A continuación se detalla el diagrama Copo de nieve.

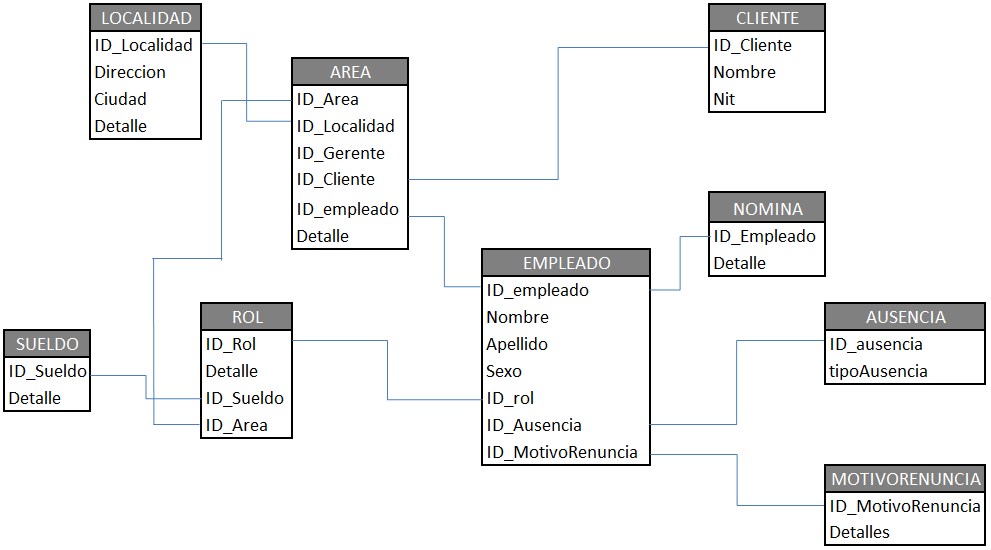
Imagen

43

.

Diagrama Copo de nieve

Fuente: Elaboración propia



Como se puede apreciar en la figura anterior, existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez estan relacionadas con una o más tablas de las otras dimensiones.

Imagen 44. Construcción del DataWarehouse



Fuente: Elaboración propia

A continuación se muestra la ilustración número 15, la cual contiene la salida del proceso de la elaboración del ETL, como aclaración los datos fueron modificados y alterados para el uso de los mismos.

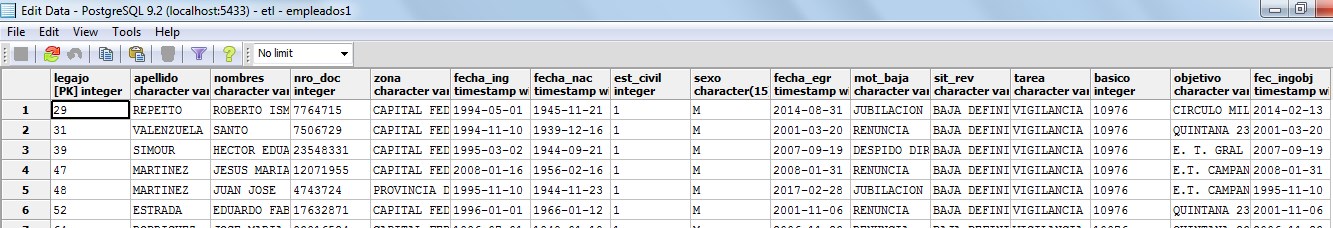
Ilustración

7

. Salida del

proceso

Fuente: Elaboración propia



La salida del archivo almacenado en el DataWarehouse es el input que recibe el algoritmo de machine Learning para poder hacer la predicción con los datos de la empresa. La siguiente fase dentro de la metodología CRISP-DM es la fase de modelado en donde se hace referencia a descubrimiento de los patrones, generación de las pruebas y diseño del prototipo funcional.

## 4.5 DISEÑO DE PRUEBAS

En este apartado se profundizará con ayuda del prototipo las variables analizadas y que se utilizaron para relevar los datos de la investigación, así como también, cual es la importancia de estos mismos para el area de recursos humanos. *En la sección 2.5.5.4 Especificación De Las Librerías Machine Learning Scikits Learn* se profundiza más sobre las herramientas usadas en la construcción del prototipo. Las variables fueron sobre la calidad de vida en las que se encuentra el individuo, como las características en el que se ubica la vivienda, el grado educativo, la seguridad social, el tiempo libre, las relaciones sociales, desempeño laboral y la gestión de los recursos humanos.

Estas mismas les dan la posibilidad y la capacidad a la gerencia de determinar el candidato a que cliente puede estar yendo a trabajar o validar si no lo contrata. La empresa debe ser un sujeto que facilite el estado de bienestar, salud y potenciar estrategias para garantizar el bienestar de los trabajadores y la calidad de vida. Como se ha mencionado anteriormente, se ha focalizado el interrogante en los vínculos que se establecen entre las personas y el mundo del trabajo, tomando como referencia un rubro de actividad específica convocante, a saber, empresa de seguridad privada.

No es el objeto de éste trabajo ofrecer el único criterio para que se contrate al postulante, todo lo contrario, brindar mejores recursos y gráficos con posibles resultados del análisis del perfil del candidato.

### 4.5.1 DETECCION DE PATRONES

El consolidado anual de contrataciones informado por la alta gerencia de la empresa Argentina de Seguridad reporta el aumento de la deserción de los empleados que ingresan y egresan por cada año de gestión laboral. El informe detalla con gran preocupación la deserción y desvinculación considerablemente en las empresas contratadas. En promedio 3 de cada 15 empleados deserta por mes. Depende de la cantidad de nuevos proyectos contratados pueden ingresan en promedio 180 empleados de lo contrario pueden egresan un promedio de 230 empleados con relación al año pasado.Para la construcción de los modelos se dispone de más de 5540 registros de empleados que ingresaron en el periodo

2006-2017 como se muestra en la Tabla 10.

Tabla 10. Empleados por periodo

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| periodo | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| empleados | 453 | 429 | 487 | 468 | 453 | 412 | 498 | 420 | 466 | 498 | 452 | 512 |

Fuente: Elaboración propia

El ausentismo, la desmotivación, la rotación, la poca focalización de objetivos de trabajo y otros factores que afectan el desarrollo social del individuo, no pueden ser disociados de sus condiciones de vida, no solo desde una intención explicativa sino como se intentará demostrar en este trabajo, también desde la consideración de los estilos de vida de dichas personas en la gestión de RRHH.

La idea de usar los algoritmos de Machine Learnig en el trabajo de investigación fue básicamente para el reconocimiento de patrones ocultos en los datos obtenidos de la empresa. El único propósito de este método fue el de clasificar grupos de categorías y conjuntos de características similares para las pruebas de predicción. Estas mismas fueron evaluadas comparándolas con un conjunto de entrenamiento (previo) llamado también set de entrenamiento. Este reconocimiento y análisis de los datos contribuyen y facilitan a tomar mejores decisiones y a realizar mejores predicciones que el ser humano no es capaz de reconocerlas a simple vista.

Los patrones que se detectaron fueron los siguientes:

* Genero poblacional: Se detectó que el género que predomina en la empresa es el masculino. Al ser una empresa que brinda seguridad privada son más los hombres que quieren trabajar en este sector.

* La edad característica en la empresa es de 31 a 45 años en promedio. También se detectó que hay un grupo pequeño de 50 a 65 años.

* En cuanto al mayor número de ausentismo por mes detectado, predominan los trabajadores con edades de 26 a 35 años en comparación con un grupo pequeño de (45 a 60)

* Zona de residencia: El lugar de residencia es el de provincia, siendo la zona sur el lugar más representativo. Históricamente la Provincia de Buenos Aires ha sido la de mayor migración poblacional, por la fuerza de su extensión geográfica.

La mayor fuerza de trabajo en la actualidad es proveniente de dichas zonas siendo así el 1er. Cordón; Avellaneda, Gral. San Martín, Lanús, Morón, Hurlingham,

Ituzaingó, San Isidro, Tres de Febrero, y Vicente López y el 2do. Cordón; Almirante Brown, Berazategui, Esteban Echeverría, Ezeiza, Florencio Varela, San

Miguel, Islas Malvinas, José C. Paz, Lomas de Zamora, La Matanza, Merlo, Quilmes, San Fernando y Tigre. Estas zonas se caracterizan por escases de recursos, pocas oportunidades laborales y precariedad en la calidad de vida, por ende deben acudir a otras zonas como capital, donde empresas como as de la presente investigación, tratan de ―satisfacer‖ dichas necesidades.

* Estado civil El estado civil que predomina es la soltería y en segundo lugar solteros con hijos.

* La jornada de trabajo que predomina es la jornada de trabajo Diurna.

* Existe una cultura saludable en los trabajadores de la empresa, realizan actividades deportivas y a su vez no consumen cigarrillo, lo que genera, mejor estado de salud con implicaciones notorias en el presentismo laboral.

* El personal femenino tiene entre 2 y 4 hijo de edades de 6 a 15 años.

* Los motivos de ausentismo son por motivos personales, los cuales abarcan desde problemas económicos, hábitos de sueño, problemas de salud entre otros. Se destacan los siguientes:
  + No tuve plata para viajar. o Me quede dormido. o Estuve enfermo. o No tenía con quien dejar mis hijos.
  + Perdí el colectivo o Utilización de PC

### 4.5.2 EMPLEADO ACTUAL

Siguiendo con el uso de la metodología CRISP-DM esta etapa es parte de la fase de evaluación del negocio, donde se interactúa con el prototipo y se evalúa las tareas asociadas a la revisión y ajuste del modelo de datos, al algoritmo y a la herramientas utilizadas como también a los aspectos de precisión y satisfacción del cliente con el resultado obtenido.

A continuación se detallan las pruebas hechas en el prototipo modulo *empleado actual,* en el mismo se evidencia una posible sugerencia para la toma de decisiones para que use el usuario de Recursos Humanos:

La primera imagen que el reclutador puede utilizar es el formulario para la carga donde puede elegir los filtros a consultar, estos son los tipos de renuncia y hacer la búsqueda para un determinado cliente ―Objetivo laboral‖.

Imagen 45. Elección de un motivo de renuncia



Fuente: Elaboración propia

En el siguiente filtro se da la elección del cliente al que el analista y/o reclutador pueden elegir para hacer una búsqueda previa a la entrevista del candidato.

Imagen

46

. Elección de un

objetivo Laboral



Fuente: Elaboración propia

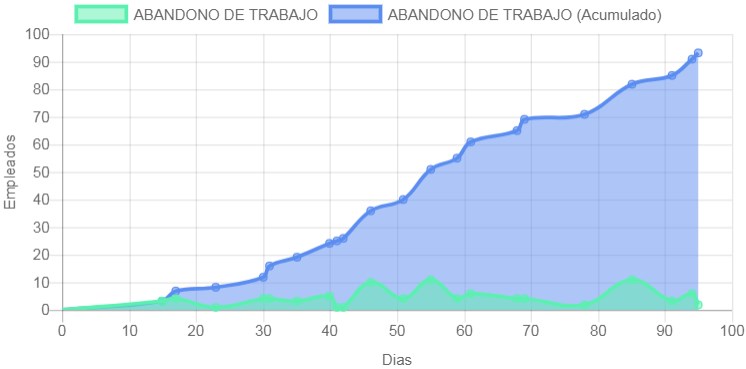
Por el perfil del candidato el reclutador al analizar la gráfica con los filtros cargados puede discernir en el tipo de contratación que puede o no ofrecer al postulante de acuerdo al resultado de la evaluación en la entrevista.

La grafica muestra un acumulado de la cantidad de los empleados asociados a este filtro, quiere decir que el analista puede ver un detalle general o individual por días.

La grafica que se muestra a continuación da el detalle del filtro usado.

Imagen 47. Empleado por días identificando el motivo de renuncia

Fuente: Elaboración propia



Se observa en la gráfica que para el filtro seleccionado *Abandono de trabajo* el resultado es que dentro de aproximadamente 50 días van a abandonar alrededor de 35 empleados activos su trabajo para el cliente *Bouchard Plaza*.

Esta ayuda visual apoya a que el reclutador piense en corto o largo plazo, ejemplo: ―esta semana abandonaría el trabajo un empleado, en dos semanas alrededor de 8 empleados‖.

Con el ejemplo anterior, se le facilita una posible toma de decisiones para que las transmita en su informe de la evaluación del candidato al jefe o supervisor. Si el negocio está urgido por la satisfacción del cliente y necesita los recursos activos para mantener el objetivo laboral lo podrá contratar o no.

La imagen 48 muestra datos básicos del empleado como el número de legajo, nombre y apellido, el motivo de renuncia asociado al filtro que se seleccionó en el primer paso, los días restantes para *(en este caso Abandono de trabajo)* y la posible sugerencia que se le ofrece al analista para que realice su respectivo análisis en función del perfil del candidato.

Imagen 48. Datos del empleado con el motivo de renuncia y con la sugerencia a tomar



Fuente: Elaboración propia

La siguiente prueba se realiza con otro filtro diferente, el cual el analista puede utilizar otro cliente o el mismo. En este caso se selecciona como tipo de renuncia el motivo *renuncia* y el cliente ubicado en puerto madero.

Imagen 49. Elección de tipo renuncia y objetivo laboral



Fuente: Elaboración propia

Se muestran los resultados obtenidos para el filtro anterior el cual muestra que en el corto plazo de 15 días se estarían acumulando un total aproximado de 20 empleados, de lo contrario puede optar por utilizar técnicas y métodos de retención para el desarrollo del trabajador en los 5-7 empleados que estarían renunciando.

Imagen 50. Gráfico empleado por días identificando el motivo de renuncia



Fuente: Elaboración propia

Las acciones generadas por el area de recursos humanos incluyen técnicas y acciones de retención de personal como un reconocimiento del trabajador, mayor capacitación y formación del trabajador, otorgar premios por presentismo y regalos empresariales entre otros.

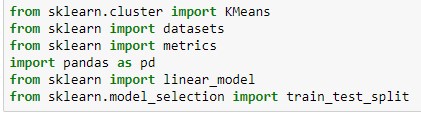
### 4.5.3 NUEVO EMPLEADO

A continuación se detalla el proceso que se utilizó para evaluar un candidato nuevo con los datos que se ingresan en el formulario que se definió para que lo usara el reclutador

*(usuarios del area de recursos humanos)*.

Primero se cargan las librerías que el algoritmo estará usando:

Imagen 51. Librerías propias de Scikit-learn



Fuente: Elaboración Propia

El funcionamiento interno del algoritmo es crear un set de datos con 1.000 registros y dos variables aleatorias con valores entre 0 y 999 X e 1 para Y. El siguiente paso será dividir los datos en validación y test mediante *train\_test\_split*:

Imagen 52. Validación *train\_test\_split*



Fuente: Elaboración Propia

El proceso del resultado es que consiste en extraer información de un conjunto de datos conocidos para clasificar nuevos datos o agrupar existentes en función de su probabilidad de pertenecer a uno u otro conjunto. Para cada nueva instancia a clasificar, se calcula la distancia con todas las instancias de entrenamiento y se seleccionan las k instancias más cercanas. A partir de un set de datos entrenados y clasificados se procede a la asignación de la clasificación para un segundo conjunto de datos de prueba.

Con la siguiente instrucción se carga en el archivo con el set de casos para entrenar el algoritmo *df = pd.read\_csv(archivo)*

El algoritmo se basa en la idea que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del set de entrenamiento más próximo. Este método supone que los vecinos más cercanos dan la mejor clasificación y esto se hace utilizando todos los atributos de las variables que se configuraron en las pruebas realizadas previas a generar el algoritmo definitivo.

Imagen 53. Matriz de asignación de campos



Fuente: Elaboración Propia

Las primeras pruebas daban un porcentaje menor para ser tenido en cuenta y evaluar el conjunto de datos de los empleados.

Imagen 54. Asignación de variable *Predict*



Fuente: Elaboración Propia

El porcentaje de la prueba resultante fue de 35% el cual es muy bajo para la aceptación correcta de la predicción.

Imagen 55. Porcentaje Obtenido



Fuente: Elaboración Propia

Posterior a estas pruebas se fueron ajustando las variables y el porcentaje fue subiendo hasta llegar a un valor acorde para las pruebas el cual oscila entre 75% a 92%. Es decir, para clasificar el nuevo dato, se basaron los siguientes casos de prueba a través de las observaciones que ya se tienen (conjunto de entrenamiento) para ver a cual se parece más

Imagen 56. Porcentaje Obtenido



Fuente: Elaboración Propia

Porcentaje óptimo.

Imagen 57. Porcentaje Obtenido



Fuente: Elaboración Propia

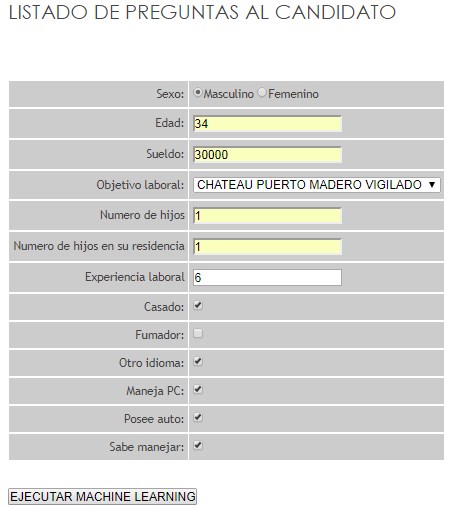
Un porcentaje mayor a 97% sobre ajusta el modelo de datos para que el algoritmo prediga los casos, quiere decir que no es muy confiable al igual que un porcentaje igual o menor a 75%.

El problema de dicha suposición es que es posible que se tengan muchos atributos irrelevantes que dominen sobre la clasificación, esto es, no elegir bien las variables de estudio y los atributos a evaluar. Adicional a este incidente de no seleccionar bien las características a evaluar se suma la lentitud que se puede llegar a dar en la clasificación cuando el algoritmo de predicción contiene mucho ruido en los datos, por ende el porcentaje superior o inferior al oprimo permitido.

Con el resultado que predijo el algoritmo, el usuario de recursos humanos podrá optar al momento de la evaluación del candidato una mejor toma de decisión. Esto es que podrá definir si es contratado bajo alguna modalidad de trabajo, o validar si en otro objetivo se ajusta más el perfil por un comportamiento similar de empleados.

El reclutador tiene otra ayuda visual mediante una tabla con las posibles sugerencias para a tomar y de esa forma pueda trasmitir al candidato en el proceso de selección. Si se ajusta a las indicaciones del negocio puede o no hacer la contratación por el tipo de necesidad. A continuación se muestran las pantallas asociadas a las pruebas realizadas con la carga de empleados en el front realizado en el framework Bottle para Python.

Imagen 58. Datos del candidato ingresados en el formulario



Fuente: Elaboración propia

Se muestra en la *imagen 47* el resultado que predijo el algoritmo indicando al usuario de recursos humanos un posible estado a futuro para el candidato a contratar.

59

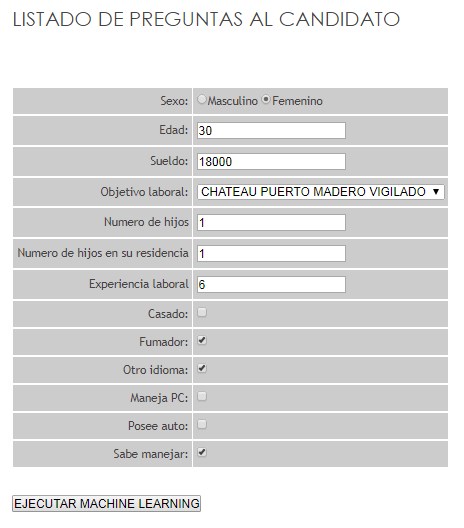


Fuente: Elaboración propia

La siguiente imagen muestra el formulario con los campos que el algoritmo toma para analizar, calcular y predecir el estado del candidato.

Imagen 60. Caso de estudio para otro candidato

Fuente: Elaboración propia



La imagen 61 muestra el resultado que el algoritmo predijo para este candidato. El mensaje que se muestra al usuario de recursos humanos no es muy alentador ya que este posible postulante con base en la configuración y datos ingresados será una renuncia para los próximos 84 días.

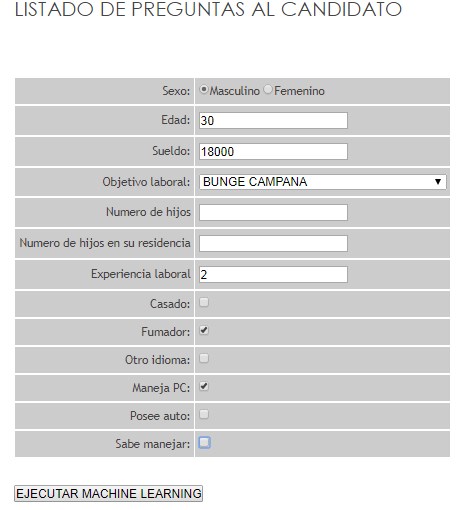
61



Fuente: Elaboración propia

A continuación se detalla un tercer caso de prueba para otro candidato con una configuración diferente en su perfil laboral y personal.

62. Caso de estudio para otro candidato



Fuente: Elaboración propia

Para este candidato el mensaje que se muestra con base en los datos ingresados será una renuncia para los próximos 23 días.

63



Fuente: Elaboración propia

El analista de recursos humanos tendrá la determinación de seguir avanzando con el proceso de selección o no de acuerdo a las necesidades del negocio, puede optar por hacer una contratación a término fijo o a una contratación de un contrato a 3 meses sin renovación laboral.

### 4.5.3 GRÁFICAS CONSOLIDADO

La última fase de la metodología CRISP-DM es la del *despliegue del modelo*. Las tareas que incluye esta fase son las de evidenciar el resultado y los reportes finales del uso de los modelos y las herramientas aplicadas para predecir correctamente un motivo de renuncia de un futuro candidato a futuro como también tareas de revisiones del proyecto a nivel de relaciones de costo y retornos de inversión.

Para tener una mejor visualización de los resultados y para optimizar la búsqueda de candidatos al usuario, el prototipo cuenta con la posibilidad de generar gráficas para el análisis a nivel macro de los objetivos laborales y de los motivos de las renuncias.

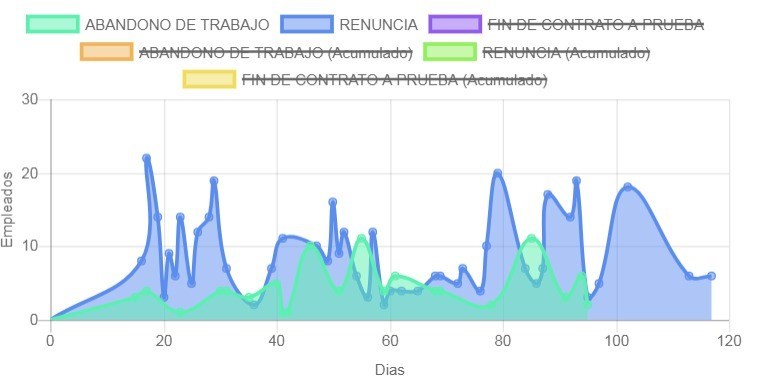
Imagen 64. Elección de tipo renuncia y objetivo laboral



Fuente: Elaboración propia

El siguiente grafico muestra el detalle de dos tipos de motivos de renuncia, estos son Abandono de trabajo y Renuncia. En el detalle del mismo hace referencia al objetivo del cliente en Chateu puerto madero en el cual tiene un alto índice de renuncia a corto plazo.

Gráfico 1. Motivos de renuncia: Abandono de trabajo y Renuncia



Fuente: Elaboración propia

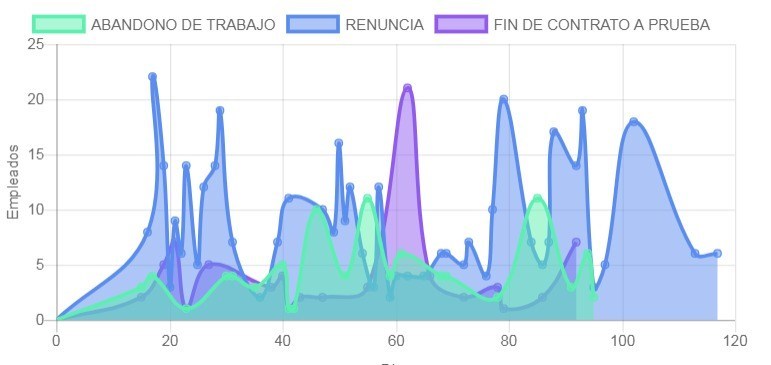
En la gráfica se muestra un total de número de empleados para cada tipo de filtro, en este caso se detalla el valor individual mostrado en la misma gráfica, quiere decir que el reclutador/analista de selección puede utilizar para ver un comportamiento a alto nivel por cada tipo de motivo de renuncia.

El siguiente grafico contiene tres filtros por motivos de renuncia que el analista puede utilizar para evidencia el comportamiento de los empleados en un corto y largo plazo en la línea de tiempo y de esa forma utilizar técnicas de retención de personal y de poder indagar en que está afectando a los trabajadores en este cliente en particular.

También puede hacer la búsqueda por diferentes clientes. En trabajos futuros el prototipo tendrá la opción de comparar los motivos en varios clientes a la vez.

Gráfico 2. Motivos de renuncia: Abandono de trabajo, Renuncia y Fin de contrato a prueba

Fuente: Elaboración propia



De la misma forma, el usuario puede seguir usando filtros de consulta para su respectivo análisis de empleados del cliente ubicado en Puerto Madero.

Gráfico 3. Motivos de renuncia: Fin de contrato a prueba

Fuente: Elaboración propia

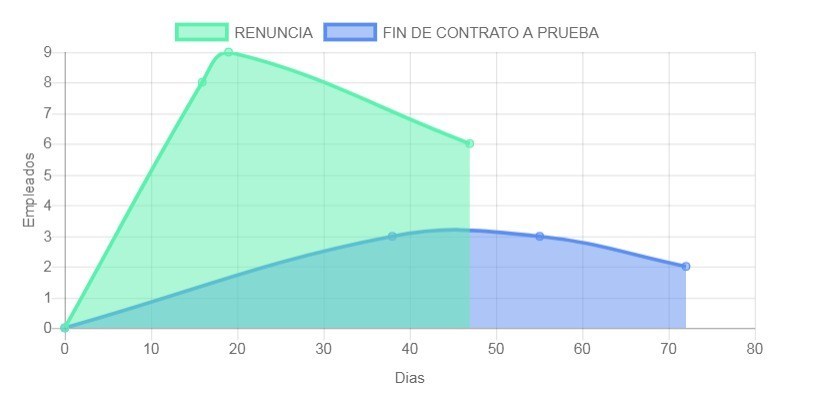


La siguiente grafica es asociada al cliente de SIEMENS, en donde se puede observar dos filtros que se seleccionan para que el analista de recursos humanos pueda verificar el estado a futuro de los empleados de dicho cliente.

De esta forma se pueden filtrar clientes por sus motivos de renuncia y detectar que técnicas o gestión se puede realizar para retener a los empleados o hacer nuevas búsquedas especificas por clientes de la empresa y prevenir insatisfacción de clientes.

Gráfico 4. Motivos de renuncia: Renuncia y Fin de contrato a prueba

Fuente: Elaboración propia



El análisis que se ha realizado, permite detallar aspectos clave para el acercamiento a tres variables, no sólo desde el punto de vista académico, sino también desde el punto de vista de la práctica empresarial. En este sentido, uno de los objetivos de la presente investigación es extraer conclusiones útiles para el ámbito de la gestión de recursos humanos con el fin de mejorar su contribución en el desempeño de los empleados y así mismo al éxito empresarial.

De esta manera, la contribución de la investigación a los ámbitos empresariales y organizacionales, debe enfocarse en valorar el concepto del capital humano de la organización en toda su amplitud, reconociendo el potencial en cada empleado y así mismo velar por el cumplimiento de un mejor desempeño.

La empresa al utilizar el prototipo está contando con una ayuda de análisis de datos con la información que tienen almacenada y de esa forma el analista de selección puede usar o no el sistema generado mediante algoritmos de machine learnig para un mejor detalle de resumen de evaluaciones de candidatos. Esto es, que el reclutador observa los gráficos que se generaron en los incisos anteriores y puede determinar si contrata o no al candidato por su perfil.

En cuanto a costos estaría previniendo futuros abandonos de empleados en diferentes clientes de la empresa y de esa forma puede llegar a implementar técnicas de retención de personal o alguna otra acción proveniente del sector de recursos humanos. El prototipo es un facilitador para la toma de decisiones, la empresa puede optar por no usarlo pero estaría perdiendo un gran potencial en el análisis de sus datos.

## 4.6 ANALISIS DE COSTOS

El costo para una empresa contratar a un recurso es casi de un 1 a 1.5 veces su sueldo si subcontrata a una *consultora* o a una *Head-hunter* para que haga todo el proceso de selección del personal, este porcentaje se paga si el recurso es contratado por el cliente de la búsqueda, de lo contrario no se realiza dicho pago. Pero, si la empresa tiene un area de Recursos Humanos *(RRHH)* el costo es asociado al sueldo que perciben sus trabajadores de esta area.

En el caso de estudio e investigación la empresa tiene conformada un area de gestión de recursos humanos y los procesos de contratación y búsqueda laboral la hacen

directamente ellos.

Se plantea una problemática de contratación con el uso y el no uso del prototipo:

Sin hacer uso del prototipo: El analista de selección entrevista y contrata a un grupo de 10 personas, de las cuales 8 candidatos tienen un perfil que van a renunciar en menos de tres meses pero él no sabe ese detalle y son contratados, la ganancia para la empresa sabiendo que el sueldo en mano es de 20.000 pesos sería aproximadamente de **$ 399.996,00** por mes y por año la empresa tendría una ganancia de **$ 2.399.976,00** haciendo los pagos respectivos al empleador. En la tabla número se detalla este valor.

Tabla 11. Ganancia por empleado

|  |  |  |  |
| --- | --- | --- | --- |
| **Ganancia por** | **empleado** |  |  |
|  |  |  |  |
| **CONCEPTO** | UNIDADES | REMUNERACION BRUTA | DESCUENTOS |
| **SUELDO** | 30 | $ 23.820,00 |  |
| **JUBILACION** |  | $ 0,11 | $ 2.620,20 |
| **LEY19032** |  | $ 0,03 | $ 600,00 |
| **OBRA SOCIAL** |  |  | $ 600,00 |
|  | | | |

**TOTAL**  $ 19.999,80

|  |  |
| --- | --- |
| **CANTIDAD DE EMPLEADOS A CONTRATAR** | 10 |
| **LA EMPRESA FACTURA EL DOBLE DEL VALOR DEL**  **RECURSO AL CLIENTE**  **GANANCIA POR AÑO TRABAJADO** | $ 199.998,00 |
| 2 |
| $ 399.996,00 |
| $ 2.399.976,00 |

De las 10 contrataciones 8 empleados renunciaran antes de los tres meses de los cuales la empresa con dos recursos ganará un total de **$ 479.995,20** por año. Se muestra y se detalla este valor a continuación:

Tabla 12. Ganancia por empleado

**Ganancia por empleado**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **CONCEPTO** | UNIDADES | REMUNERACION BRUTA | DESCUENTOS |
| **SUELDO** | 30,00 | $ 23.820,00 |  |
| **JUBILACION** |  | $ 0,11 | $  2.620,20 |
| **LEY19032** |  | $ 0,03 | $  600,00 |
| **OBRA SOCIAL** |  |  | $  600,00 |
|  |  |  |  |
|  | **TOT** | **AL** | $  19.999,80 |
|  |  |  |  |

**CANTIDAD DE EMPLEADOS A CONTRATAR** 2

$

39.999,60

|  |  |  |
| --- | --- | --- |
| **LA EMPRESA FACTURA EL DOBLE DEL VALOR DEL**  **RECURSO AL CLIENTE** |  | 2 |
| $  79.999,20 |  |
|  |  |  |

**GANANCIA POR AÑO TRABAJADO** 12

$

479.995,2

0

Quiere decir que la empresa estaría dejando de ganar por año un valor de $ **1.919.980,80**.

Por los recursos que renunciaron.

Usando el prototipo el analista de selección podrá hacer uso de un análisis del perfil del candidato y observar en las gráficas para qué cliente podría aplicar mejor su perfil. Está en el analista de selección tomar la decisión que mejor se ajuste a su búsqueda, esto quiere decir que solamente el analista sabe si lo contrata o no por alguna directriz de la gerencia. En otras palabras, puede ser que la empresa tenga alguna ganancia contratando a un recurso que renunciará en tres meses y solventa esa posición con otro recurso de otro cliente. Acá es donde juega un papel importante el analista de selección y ―mover a los recursos‖ en donde sabe que el índice de rotación es bajo o alto.

El conocimiento permite tomar decisiones más acertadas. El area de gerencia humana se ve directamente beneficiado por este tipo de análisis, ya que tener visibilidad de los datos de los colaboradores permite tomar decisiones más atinadas. Al utilizar machine learning se está yendo un paso más allá, porque se podrá no solo entender que pasó sino también lo que va a suceder y determinar qué alternativas o técnicas de retención de personal se pueden implementar al respecto.

# 5. RESULTADOS Y CONCLUSIONES

De acuerdo a la conceptualización de esta investigación y dando respuesta al interrogante planteado se puede decir que la gestión de las empresas, y de cada una de sus áreas en especial la de Recursos Humanos, debe velar por la construcción de un vínculo virtuoso, humano y de retención del talento.

Las organizaciones van comprendiendo, que tienen que captar aliados y no enemigos, porque de otro modo no les resulta posible ganar las batallas de competitividad y expansión del mercado y en muchos casos se pone en juego la supervivencia.

Muchas empresas buscan trascender la concepción básica del vínculo trabajador, empleador mediante políticas de compensaciones, beneficios o recompensas, estrategias de motivación o implementando liderazgo.

Sin embargo persisten políticas que apuntan a disciplinar a la persona más que a situar la problemática en un enfoque basado en su condición social o en la consideración de estrategias de afrontamiento de las demandas de diferentes contextos que deben satisfacerse para que una persona pueda cumplir con un régimen específico de trabajo, las cuales en ocasiones fracasan.

En suma, la dirección de las empresas y, en particular, el área de recursos humanos, debería valorar al recurso humano en una dimensión más completa –intelectual, social y afectiva- con el objetivo de aplicar las prácticas y políticas que puedan potenciar cada una de estas dimensiones del capital humano de sus empleados. Todas ellas son necesarias para maximizar el valor que puedan aportar los empleados a la organización a través de un capital humano multidimensional.

A continuación se detallan los resultados de la investigación.

## 5.1 RESULTADOS

Una vez probados y evaluados los modelos de minería de datos, se comprueba que los factores principales estudiados influyen en el desempeño laboral generan que el empleado deserte. Los resultados fueron evaluados por la comprensión e interpretación de los resultados del modelo, así como el impacto de los resultados de minería de datos para los objetivos del negocio. Las variables que se consideraron en la utilización de la estructura de los modelos de minería de datos fueron consideradas y evaluadas para el comportamiento adecuado de cada modelo.

La vulnerabilidad incluye tres tipos de riesgos relacionados a, exposición a la crisis, a una falta de capacidad para afrontarlas y el riesgo de sufrir consecuencias graves a causa de ellas. Estos tipos de riesgos, manifiestan la vulnerabilidad que puede atravesar una persona y evidentemente las estrategias que se deben generar para poder sortearlos. Las estrategias permiten que el individuo desarrolle capacidades y habilidades para entender la situación y poder entablar acciones correctivas a las mismas.

La vulnerabilidad laboral, pone en riesgo a la persona para afianzar las capacidades que tiene para desarrollarse y cumplir metas de realización pasando por situaciones como las siguientes:

* Sentirse excluido por su condición de género o raza
* Atravesar por dificultades de rechazo debido a sus antecedentes penales.
* Dificultarse la comprensión de tareas y ejecución de las mismas debido a la falta de experiencia en empresas
* Ser sostén de familia y tener que reducir su jornada de trabajo así como su ingreso dada la responsabilidad con los mismos.
* Poseer problemas de salud y no tener acceso a cobertura médica
* No poseer disponibilidad y acceso a diferentes medios de transporte para la movilización al trabajo

La gestión de Recursos Humanos y las condiciones de vida determinan el desempeño laboral en los trabajadores de empresas de servicios. Se puede decir que los estudios exploratorios ayudan a obtener información para realizar una investigación más completa de un contexto particular, investigar nuevos problemas, identificar conceptos o variables promisorias, establecer prioridades para investigaciones futuras o sugerir afirmaciones y postulados.

## 5.2 CONCLUSIONES

Los procesos de minería de datos están orientados al desarrollo de métodos que exploran los datos almacenados en las bases de datos de las empresas así como el uso de las técnicas para transformar dichos datos en información y entender mejor el proceso de aprendizaje y conocimiento de los mismos. El motivo de la investigación de tesis es dar inicio y visibilidad a las aplicaciones de sistemas basados en modelos de minería de datos orientados a los sectores laborales para analizar y evaluar los factores que influyen principalmente en la deserción de los trabajadores. Los analistas de recursos humanos de la empresa Argentina de Seguridad podrán usar la metodología propuesta por este trabajo para identificar y establecer procedimientos que permitan detectar en forma temprana las variables relevantes con el objeto de mejorar los índices de deserción laboral. Durante el desarrollo del trabajo se reconoce que fueron alcanzados los objetivos propuestos.

Se estudiaron diferentes técnicas para desarrollar modelos de predicción utilizando minería de datos, tales como árboles de decisión, redes neuronales y técnicas de clasificación, como clustering, k-mediana con el objetivo de aplicar las herramientas adecuadas, probando en si los modelos para clasificar a los empleados con mayor riesgo de deserción. Las empresas, se enfrentan a un gran desafío: comprender los diversos entornos por los que su personal atraviesa y así mismo generar política de fortalecimiento para su superación diaria.

Los encargados de reclutamiento en el area de recursos humanos deben hacer una búsqueda explicita de los requisitos de competencia que se exige al personal a contratar para ocupar un determinado puesto de trabajo y definido por la empresa solicitante. Con la ayuda del prototipo se pueden ayudar en la correcta toma de contratación ya que se predice por el tipo de rol del candidato su estado a futuro y su posible situación laboral dentro de la empresa.

La estructura de una organización debe ser resultado de la selección de elementos, teniendo en consideración la búsqueda de consistencia interna (procesos internos) y externa (medio ambiente y contexto).La rotación de personal es necesaria hasta cierto nivel para ir mejorando la productividad de la organización al ir buscando debilidades y promover cambios. Sin embargo, es inevitable también, cierta cantidad de movilidad por motivos de enfermedades, accidentes, envejecimiento, muerte y otras razones personales que producen la deserción de un puesto de trabajo. Las empresas deberían insistir en el desarrollo de su capacidad de innovación como vía para lograr el éxito empresarial y comprender el papel del capital humano en dicho desarrollo, contemplando la unidad básica sin la cual no pueden lograrlo: el trabajador.

En algunas situaciones dados los valores obtenidos, no se pudo validar positivamente los modelos de árbol de decisión y cluster k-mediana, dado que su capacidad de predicción es menor que la estimación aleatoria, es decir se necesitaba de un mayor número de datos para poder aplicarlos correctamente. El análisis previo a la investigación que se desarrolló permitió apreciar la importancia que tiene el proceso de recopilación de datos, abarcando las fases de análisis y preparación de los datos asociado a la metodología CRISP-DM. El aprendizaje automático es una técnica de análisis de datos que enseña a los ordenadores a hacer lo que resulta natural para las personas y los animales: aprender de la experiencia. Estos a su vez emplean métodos de cálculo para ―aprender‖ información directamente de los datos. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje.

## 5.3 TRABAJO FUTURO

* Reconocimiento por la foto del empleado que se puede cargar en la base de datos y analizarla mediante algoritmos *TensorFlow*, redes neuronales, Algoritmos Supervisados.
* Implementar el algoritmo en distintos rubros, donde la modificación de variables sea mínimo.

# 6. BIBLIOGRAFÍA

Agrawal, R., Imielinski, T., Swami, A.: Database Mining: A performance Perspective.

IEEE Trans. Knowledge and Data Engineering, vol. 5, 6 (1993b) 914-925.

Agrawal, R., Srikant, R.: Fast Algorithms for mining association rules in large databases. Proc. of 20th Int. Conference on Very Large Databases, Santiago de Chile (1994)

487-489.

Alaa el-Halees (2009), "Mining Students Data To Analyze Learning Behavior: A Case

Study," Gaza, 2009.

Base de datos orientada a objetos. (2018, Febrero 27). En *Wikipedia, la enciclopedia libre*. Recuperado a partir de [https://es.wikipedia.org/w/index.php?title=Base\_de\_datos\_orientada\_a\_objetos&o ldid=105876460](https://es.wikipedia.org/w/index.php?title=Base_de_datos_orientada_a_objetos&oldid=105876460)

Berry M. y Linoff G. (2004). Data Mining Techniques: for marketing, sales, and customer relationship management. 2da edición por Wiley Publishing, Inc., pp.

643, ISBN: 0-471-47064-3, Indiana, Estados Unidos.

Brewer, D. E. A. (2000). Towards Robust Distributed Systems, 12.

Chapman et al. (2000) CRISP-DM Methodology, Download Scientific Diagram. (s. f.).

Recuperado 26 de junio de 2018, a partir de

https://www.researchgate.net/figure/CRISP-DM-Methodology-Chapman-et-al-

2000\_fig1\_313492466

Charles W. Bachman (1992), Business Software Innovator, Dies at 92 - The New York Times. (s. f.). Recuperado 12 de febrero de 2018, a partir de

<https://www.nytimes.com/2017/07/16/technology/charles-w-bachman-dies.html>

CM Bishop, Pattern Recognition and Machine Learning, Fig 9.1, p426

Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks". In:

Communications of the ACM 13 (6): 377-387.

Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business (1st edición). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42, more. ISBN

978-0-9893086-0-1.

Eckerson, Wayne (2007), extending the Value of Your Data arehousing Investment, The Data Warehouse Institute

Ecuación del centroide 1886f2c69775746ac7b6c1cdd88c53c676839015.png (156×50). (s. f.). Recuperado 24 de junio de 2018, a partir de [http://scikithttp://scikit-learn.org/stable/\_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.pnglearn.org/stable/\_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.pn g](http://scikit-learn.org/stable/_images/math/1886f2c69775746ac7b6c1cdd88c53c676839015.png)

E. F. Codd (1972). Communications of the ACM, volume=13, issue=6, pages=377-387

Fabian Pedregosa; Gaël Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand

Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent

Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu

Perrot; Édouard Duchesnay (2011). "Scikit-learn: Aprendizaje automático en

Python" . Journal of Machine Learning Research . 12 : 2825-2830.

Fayyad, Usama (1996). ―Advances in Knowledge Discovery and Data Mining‖. MIT

Press.

Geng L. y Hamilton H. (2007). Choosing the Right Lens: Finding what is Interesting in Data Mining. Studies in Computational Intelligence (SCI), 43, 3-24. Editado por

Guillet F. y Hamilton H. Quality Measures in Data Neevia docConverter 5.1

Mining, editado por Springer-Verlag, pp. 313, ISBN: 3-540-44911-6, Heidelberg,

Alemania.

Gironés, Jordi, et al. Minería de datos: modelos y algoritmos, Editorial UOC, 2017. ProQuest Ebook Central, [http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=5045 398](http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=5045398)

Gironés, J., Casas, J., & Minguillón, J. (2017). Minería de datos: modelos y algoritmos.

Barcelona, SPAIN: Editorial UOC. Recuperado a partir de

[http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=5045 398](http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=5045398)

Gomez-Flechoso, A.J.(1998), Inducción de Conocimiento con incertidumbre en Bases de

Datos Relacionales Borrosas.

Hair J., Anderson R., Tatham R. y Black W. (1995). Multivariate data analysis. Englewood Cliffs, 4ta edition por Prentice Hall, pp. 500, ISBN:0023490209,

Estados Unidos.

Halper, Fern (2011), «The Top 5 Trends in Predictive Analytics», Information

Management

Han, J.; Kamber M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann

Publishers, USA.

Hand D., Mannila H. y Smyth P. (2001). Principles of Data Mining, editado por The Massachusetts Institute of Technology Press., pp. 546, ISBN: 0-262-08290- x,

Massachusetts, Estados Unidos.

Hernández (2004). J., Ramírez M. J. y Ferri C. (2004). Introducción a la Minería de Datos. Pearson Educación. Editorial Pearson Prentice Hall, pp. 680, ISBN: 84-

205-4091-9, Madrid, España.

Hernández Orallo, J. (2004). Introducción a la Minería de Datos: Pearson.

Iruela, J. (2016, enero 19). Los gestores de bases de datos más usados. Recuperado 11 de julio de 2018, de https://revistadigital.inesem.es/informatica-y-tics/los-gestoresde-bases-de-datos-mas-usados/

Jain A., Murty M. y Flynn P. (1999). Data Clustering: A Review. ACM Computing

Surveys, 31, 3, 264-323, ISBN: 0360-0300, New York, Estados Unidos.

Jill Dyché (2000). e-Data: turning data into information with data warehousing. Addison-

Wesley. p.323.

Kimball, Ralph; Margy Ross (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (2nd ed.). Wiley. ISBN 0-471-20024-7.

Kimball (2013). Ralph; Margy Ross (2013). The Data Warehouse Toolkit: The Definitive

Guide to Dimensional Modeling (3rd ed.). Wiley. ISBN 978-1-118-53080-1.

Klösgen W. y Zytkow J. M. (2002). Handbook of Data Mining and Knowledge

Discovery. Oxford University Press, Inc., pp. 980, ISBN: 0-19-511831-6, New

York, Estados Unidos.

Larose D. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John

Wiley & Sons, Inc., pp. 222, ISBN: 0-471-66657-2, New Jersey, Estados Unidos.

MacLennan, Jamie (2012), 5 Myths about Predictive Analytics, The Data Warehouse

Institute

McKinney, W. (s. f.). Python for Data Analysis. Recuperado a partir de http://shop.oreilly.com/product/0636920050896.do

Milligan G. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 3, 325-342.

Mimer. (s. f.). Recuperado 02 de Noviembre 2017, a partir de <https://www.mimer.com/>

Moine, I. J. M., Haedo, D. A. S., & Gordillo, D. S. (s. f.). Estudio comparativo de metodologías para minería de datos, 4.

Muñoz D. y Rodrigo F. (2002). Aplicación del análisis cluster para el estudio de la relación Nao-Precipitaciones de invierno en el sur de la Península Ibérica.

Departamento de Física Aplicada, Universidad de Almería, pp. 283-292, España.

Nancy Lynch and Seth Gilbert, ―Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services‖, ACM SIGACT News, Volume 33 Issue 2 (2002), pg. 51-59

NoSQL Relational Database Management System: Home Page. (s. f.). Recuperado 24 de junio de 2018, a partir de [http://www.strozzi.it/cgihttp://www.strozzi.it/cgi-bin/CSA/tw7/I/en\_US/nosql/Home Pagebin/CSA/tw7/I/en\_US/nosql/Home%20Page](http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/nosql/Home%20Page)

Nyce, Charles (2007), Predictive Analytics White Paper, American Institute for

Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

OLAP. (2018, marzo 20). En *Wikipedia, la enciclopedia libre*. Recuperado a partir de <https://es.wikipedia.org/w/index.php?title=OLAP&oldid=106372242>

Palma, Méndez, José Tomás, and Morales, Roque Marín. Inteligencia artificial: métodos, técnicas y aplicaciones, McGraw-Hill España, 2008. ProQuest Ebook Central, [http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=3194 97](http://ebookcentral.proquest.com/lib/bibliouniminutosp/detail.action?docID=319497)

Petrovic, D. (2010). SQL Server 2008 Manual de referencia. Recuperado el 3 de Marzo de 2018 Alemania: Mc Graw Hill.

Raimason (2018). Database Management System - Product Comparison. (s. f.). Recuperado 24 de febrero de 2018, a partir de [https://raima.com/producthttps://raima.com/product-comparison/comparison/](https://raima.com/product-comparison/)

Ralph Kimball; Margy Ross; Warren Thornthwaite; Joy Mundy (January 10, 2008). The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses (Second ed.). Wiley. ISBN 978-0-470-14977-5.

Simon Haykin(1999), Neural Networks. New York: Macmillan College (IEEE Press

Book), 1994. Jain A., Murty M. y Flynn P. (1999). Data Clustering: A Review.

ACM Computing Surveys, 31, 3, 264-323, ISBN: 0360-0300, New York, Estados

Unidos.

Sistema de gestión de bases de datos. (2018, mayo 16). En *Wikipedia, la enciclopedia libre*. Recuperado a partir de [https://es.wikipedia.org/w/index.php?title=Sistema\_de\_gesti%C3%B3n\_de\_bases \_de\_datos&oldid=107869791](https://es.wikipedia.org/w/index.php?title=Sistema_de_gesti%C3%B3n_de_bases_de_datos&oldid=107869791)

Sumathi S. y Sivanandam S. (2006). Introduction to Data Mining and its App.lications.

Studies in Computational Intelligence, 29, editado por Springer-Verlag, pp. 828,

ISBN: 3-540-34350-4, Heidelberg, Alemania.

Torres-Jiménez, J. (2011), Breve Introducción a las Bodegas de Datos.

[http://www.tamps.cinvestav.mx/~jtj/courses/dbs/slides/Bodegas%20de%20datos.](http://www.tamps.cinvestav.mx/~jtj/courses/dbs/slides/Bodegas%20de%20datos.pdf)

[pdf](http://www.tamps.cinvestav.mx/~jtj/courses/dbs/slides/Bodegas%20de%20datos.pdf)

Usama Fayyad y Evangelos Simoudis. (1997). Data Mining and Knowledge Discovery in Databases

Vazirgiannis M., Halkini M. y Gunopulos D. (2003). Uncertainty Handling and Quality

Assessment in Data Mining. Advanced Information and Knowledge Processing,

editado por Springer-Verlag, pp. 226, ISBN: 1-85233-655-2, Heidelberg,

Alemania.

wal, R., Imielinski, T., Swami, A.: Mining associations between sets of items in large databases. Proc. of ACM SIGMOD Int. Conference on Management of Data

Washinton, D.C. (1993a) 207-216.

Wagsta, K., Cardie, C., Rogers, S., & Schroedl, S. (s. f.). Constrained K-means

Clustering with Background Knowledge, 8.

Whitehorn, Mark (2010). "Inmon vs. Kimball data warehousing: the debate over DW architecture". ComputerWeekly.com. Archived from the original on 2015-10-03.

Retrieved 2015-12-11.

Witten I. y Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2da ed. por Morgan Kaufmann Series in Data Management Systems, pp. 525, ISBN: 0-12-088407-0, Estados Unidos.

Zhu, X. y I. Davidson. (2007), Knowledge Discovery and Data Mining: Challenges and

Realities. IGI Global.

1. Glosario básico de términos utilizados en Inteligencia Artificial. (2017, octubre 2). Recuperado 15 de julio de 2018, de http://observatorio-ia.com/glosario-basico-de-terminos-utilizados-en-ia [↑](#footnote-ref-1)
2. ANSI/X3/SPARC Study Group on Data Base Management Systems: (1975), Interim Report. FDT, ACM SIGMOD bulletin. Volume 7, No. 2. [↑](#footnote-ref-2)
3. Elizabeth N Fong, David K Jefferson, Reference models for standardization, In Computer Standards & Interfaces, Volume 5, Issue 2, 1986, Pages 9398, ISSN 0920-5489[, https://doi.org/10.1016/0920-5489(86)90075-9.](https://doi.org/10.1016/0920-5489(86)90075-9)

   [↑](#footnote-ref-3)
4. ANSI/X3/SPARC Database System Study Group. Database Architecture Framework Task Group. (1985). Reference model for DBMS standardization.

   [Washington, D.C.?] : [Springfield, VA : U.S. Dept. of Commerce, National Bureau of Standards ; National Technical Information Service, distributor [↑](#footnote-ref-4)
5. Tomado de: Noticias, artículos y guías sobre estrategia y operaciones de TI en las empresas. - SearchDataCenter.com/es. (s. f.). Recuperado 24 de junio de 2018, a partir d[e https://searchdatacenter.techtarget.com/es](https://searchdatacenter.techtarget.com/es)  [↑](#footnote-ref-5)
6. About Bill : William H. Inmon, "The Father of Data Warehousing". Inmon Consulting Services 2007 [↑](#footnote-ref-6)
7. Tomado de: Proceso y herramientas ETL. (s. f.). Recuperado 24 de junio de 2018, a partir d[e http://etl-tools.info/es/bi/proceso\_etl.htm](http://etl-tools.info/es/bi/proceso_etl.htm)  [↑](#footnote-ref-7)
8. Tomado de: Dimensional modeling. (2018, mayo 3). En *Wikipedia*. Recuperado a partir de <https://en.wikipedia.org/w/index.php?title=Dimensional_modeling&oldid=839460441> [↑](#footnote-ref-8)
9. Recuperado a partir de: ETL En *Wikipedia, la enciclopedia libre*. <https://es.wikipedia.org/w/index.php?title=Extract,_transform_and_load&oldid=107626881> [↑](#footnote-ref-9)
10. ETL What it is and why it matters[, https://www.sas.com/en\_us/insights/data-management/what-is-etl.html](https://www.sas.com/en_us/insights/data-management/what-is-etl.html)  [↑](#footnote-ref-10)
11. Posted by Techroba on December 9, 2015 at 4:00am, & Blog, V. (s. f.). 10 Open Source ETL Tools. Recuperado 24 de junio de 2018, a partir de <https://www.datasciencecentral.com/profiles/blogs/10-open-source-etl-tools> [↑](#footnote-ref-11)
12. Recuperado a partir de: Sistemas, P. por S. | I. y. (s. f.). Herramientas para implementar un Data Warehouse.

    <http://softpei.blogspot.com/2016/08/herramientas-para-implementar-un-data.html> [↑](#footnote-ref-12)
13. ¿Qué es PMML? (2011, abril 6). Recuperado 24 de junio de 2018, a partir d[e http://www.ibm.com/developerworks/ssa/industry/library/indhttp://www.ibm.com/developerworks/ssa/industry/library/ind-PMML1/index.htmlPMML1/index.html](http://www.ibm.com/developerworks/ssa/industry/library/ind-PMML1/index.html)  [↑](#footnote-ref-13)
14. Archiveddocs. (s. f.). Algoritmo de árboles de decisión de Microsoft. Recuperado 24 de junio de 2018, a partir d[e https://docs.microsoft.com/eshttps://docs.microsoft.com/es-es/previous-versions/sql/sql-server-2008-r2/ms175312(v%3dsql.105)es/previous-versions/sql/sql-server-2008-r2/ms175312(v%3dsql.105)](https://docs.microsoft.com/es-es/previous-versions/sql/sql-server-2008-r2/ms175312(v%3dsql.105))  [↑](#footnote-ref-14)
15. Mike Gualtieri. «The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015». Forrester. [↑](#footnote-ref-15)
16. Python Software Foundation. (2013, marzo 9). En Wikipedia, la enciclopedia libre. Recuperado a partir de

    https://es.wikipedia.org/w/index.php?title=Python\_Software\_Foundation&oldid=64570622 [↑](#footnote-ref-16)
17. scikit-learn: machine learning in Python — scikit-learn 0.19.1 documentation. (s. f.). Recuperado 19 de abril de 2018, a partir de http://scikitlearn.org/stable/ [↑](#footnote-ref-17)
18. History of SciPy -. (2007, septiembre 27). Recuperado 20 de abril de 2018, a partir de ttps://web.archive.org/web/20070927015018/http://www.scipy.org/History\_of\_SciPy/ [↑](#footnote-ref-18)
19. Package overview — pandas 0.22.0 documentation. (s. f.). Recuperado 20 de abril de 2018, a partir de https://pandas.pydata.org/pandasdocs/stable/overview.html#license [↑](#footnote-ref-19)
20. conda-forge | community driven packaging for conda. (s. f.). Recuperado 20 de abril de 2018, a partir de https://conda-forge.org/ [↑](#footnote-ref-20)
21. Gartner, Inc., ―Magic Quadrant for Data Integration Tools,‖ Mark A. Beyer, Eric Thoo, Mei Yang Selvage, Ehtisham Zaidi, Jacki Williamson, Agosto 2017. [↑](#footnote-ref-21)
22. Azevedo, A. y Santos, MF KDD, SEMMA y CRISP-DM: una visión paralela. En Actas de la Conferencia Europea de IADIS sobre Minería de Datos 2008, pp 182-185. Archivado el 9 de enero de 2013 en Wayback Machine. [↑](#footnote-ref-22)
23. Sitio web de SAS Enterprise Miner Archivado el 8 de marzo de 2012 en Wayback Machine [↑](#footnote-ref-23)
24. CRISP-DM by Smart Vision Europe. (s. f.). Recuperado 24 de junio de 2018, a partir d[e http://crisp-dm.eu/](http://crisp-dm.eu/)  [↑](#footnote-ref-24)
25. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System [↑](#footnote-ref-25)