Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN

Pablo Fernando Mendoza Zepeda

Asesorado por el Ing. David Armando Chang Ovando

Guatemala, agosto de 2013

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

PABLO FERNANDO MENDOZA ZEPEDA
ASESORADO POR EL ING. DAVID ARMANDO CHANG OVANDO

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, AGOSTO DE 2013

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Ing. Alfredo Enrique Beber Aceituno
VOCAL II	Ing. Pedro Antonio Aguilar Polanco
VOCAL III	Inga. Elvia Miriam Ruballos Samayoa
VOCAL IV	Br. Walter Rafael Véliz Muñoz
VOCAL V	Br. Sergio Alejandro Donis Soto
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADOR	Ing. Edgar Estuardo Santos Sutuj
EXAMINADORA	Inga. Virginia Victoria Tala Ayerdi
EXAMINADOR	Ing. Edgar Roberto Pinillos Montenegro
SECRETARIA	Inga. Marcia Ivónne Véliz Vargas

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha 20 de marzo de 2013.

Pablo Fernando Mendoza Zepeda

Guatemala, 30 de Julio 2013

Ingeniero
Carlos Azurdia
Coordinador
Comisión de Aprobación y Revisión de Tesis
Carrera de Ingeniería en Ciencias y Sistemas
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

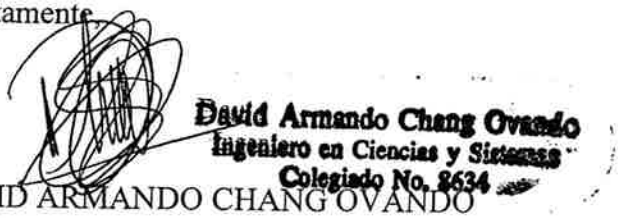
Estimado Ingeniero Azurdia:

Por medio de la presente me permito hacer de su conocimiento que he procedido a revisar el trabajo de Tesis titulado "MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN", elaborado por el estudiante PABLO FERNANDO MENDOZA ZEPEDA.

En mi calidad de asesor, he analizado el contenido, así como las conclusiones y recomendaciones expuestas. Después de haber hecho las modificaciones pertinentes, dejo constancia de mi aprobación considerando que el trabajo cumple con los objetivos propuestos para su desarrollo.

Sin otro particular, me suscribo.

Atentamente



DAVID ARMANDO CHANG OVANDO
Ingeniero En Ciencias Y Sistemas
Colegiado No. 8634

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 07 de Agosto de 2013

Ingeniero
Marlon Antonio Pérez Turk
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Pérez:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **PABLO FERNANDO MENDOZA ZEPEDA** carné **2000-10624,** titulado: **"MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN"**, y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

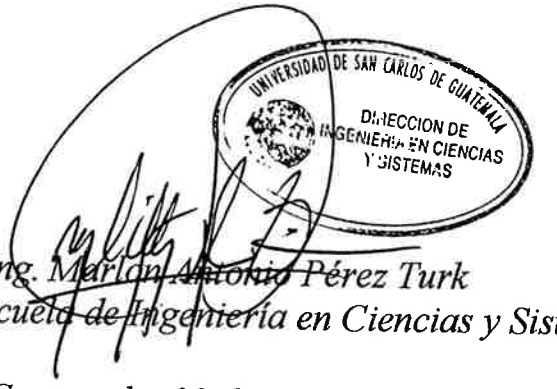
UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE CIENCIAS Y SISTEMAS
TEL: 24767644

ESCUELA DE CIENCIAS Y SISTEMAS

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación ***"MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN"****, realizado por el estudiante PABLO FERNANDO MENDOZA ZEPEDA, aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"



Ing. Marlon Antonio Pérez Turk
Director, Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 30 de agosto 2013

Universidad de San Carlos
de Guatemala

Facultad de Ingeniería
Decanato

DTG. 604.2013

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al Trabajo de Graduación titulado: **MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN,** presentado por el estudiante universitario: **Pablo Fernando Mendoza Zepeda,** autoriza la impresión del mismo.

IMPRÍMASE:

Ing. Murphy Olympo Paiz Recinos
Decano

Guatemala, 30 de agosto de 2013

/gdech

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO
FACULTAD DE INGENIERIA

ACTO QUE DEDICO A:

Mis padres	Iván Mendoza y Lesbia Zepeda. Pilares fundamentales de mi vida, les agradezco su amor y apoyo incondicional. Este título es para ustedes.
Mi esposa	Marla Menchú. Es una bendición que estés a mi lado todos los días, en especial hoy para que celebremos este logro. Te amo.
Mi hijo	Fernando Andrés. Le diste otro sentido a mi vida. Eres mi mayor inspiración.
Mis hermanos	Lucía, Iván y Sofía Mendoza. Mis queridos cómplices de este emprendimiento por la vida, he aquí un logro más para nuestro equipo. Los quiero.
Mis abuelos	Pablo Mendoza y Estela Perdomo (q.e.p.d.). Su amor y consejos siguen en mi corazón. Un saludo hasta el cielo, sé que desde allá nos protegen y comparten conmigo este momento.
Mis abuelos	Alberto Zepeda y Margarita Guzmán. Me enorgullece poder compartir este triunfo con ustedes, son parte esencial de mi vida.

AGRADECIMIENTOS A:

La Universidad de San Carlos de Guatemala	La casa de estudios que permitió cumplir con mi formación académica y sentó las bases para mi futura carrera como profesional.
Facultad de Ingeniería	Lugar que me dotó de enseñanzas y conocimientos para ejercer como profesional en el área de tecnologías de información.
Mi asesor Ing. David Chang	Por la orientación y aporte de conocimiento profesional brindado para el desarrollo de este trabajo.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Algoritmo Conjunto de reglas o instrucciones, ordenadas y finitas que permiten realizar una actividad mediante la ejecución de pasos sucesivos.

Clima organizacional Se refiere a las características del medio ambiente de trabajo, las cuales son percibidas directa o indirectamente por los trabajadores.

Cluster Es un grupo de datos que puede ser analizados en razón a la relación de las variables o atributos.

CRISP-DM Acrónimo para Cross Industry Standard Process for Data Mining, es una metodología propuesta por la industria de minería de datos que incluye descripciones de las fases típicas de un proyecto de minería de datos, las tareas concernientes a cada fase, su explicación y la relación entre estas tareas.

Escala Likert Es una escala psicométrica de uso más amplio en encuestas para la investigación, principalmente en ciencias sociales. Al responder a una pregunta de un cuestionario elaborado con la técnica de Likert, se especifica el nivel de acuerdo o desacuerdo con una declaración.

Escala nominal	Son escalas utilizadas para clasificar objetos o fenómenos, según ciertas características, tipologías o nombres, dándoles una denominación o símbolo, sin que implique ninguna relación de orden, distancia o proporción entre los objetos o fenómeno.
ERP	Acrónimo para Enterprise Resource Planning, se refiere a un sistema de información gerencial que integra y maneja el control de actividades de negocios como ventas, entregas, pagos, producción, administración de inventarios, calidad de administración y la administración de recursos humanos.
Estratek	Empresa guatemalteca que ofrece herramientas y métodos cuantitativos para la gestión del talento, como clima organizacional y evaluación del desempeño.
Minería de datos o *Data Mining*	Proceso mediante el cual se extrae conocimiento implícito, previamente desconocido y potencialmente útil, a partir de grandes volúmenes de datos; utilizando técnicas de inteligencia artificial, análisis matemático y estadístico para descubrir patrones y tendencias en los datos.
PostgreSQL	Es un sistema de gestión de base de datos relacional orientada a objetos y libre, publicado bajo la licencia BSD.

R	Es un software de código abierto que posee un entorno en el que se pueden ejecutar técnicas estadísticas, tanto clásicas como modernas.
Software	Es el conjunto de los programas de cómputo, procedimientos, reglas, documentación y datos asociados, que forman parte de las operaciones de un sistema de computación.
Software de Código Abierto	Es el software cuyo código fuente y otros derechos que normalmente son exclusivos para quienes poseen los derechos de autor, son publicados bajo una licencia de software que forman parte del dominio público.

RESUMEN

El presente trabajo documenta un proceso completo de minería de datos aplicado a una base de datos de encuestas de clima organizacional, realizadas en empresas guatemaltecas entre los años 2010 y 2012. Esto con el objetivo de descubrir patrones y modelos de información, que permitan a las empresas entender de mejor manera cuáles son las necesidades y factores que un colaborador percibe como elementos de un buen clima organizacional.

En la primera parte del capítulo 1 se presenta un marco teórico sobre el clima organizacional, su definición, importancia, metodología utilizada en las evaluaciones y el instrumento de medición. Finalmente se explica la situación actual de las herramientas que se ofrecen en el mercado, para evaluar el clima organizacional de una empresa.

En la segunda parte del marco teórico se realiza una revisión bibliográfica de la minería de datos: definición, ventajas, metodología, tipos de modelos de minería de datos y algoritmos relacionados.

En el segundo capítulo se explica la metodología de trabajo utilizada, una explicación de por qué se eligió el software *R* para ejecutar los algoritmos y finalmente se documenta todo el proceso de minería de datos, cuyo desarrollo se apegó a una metodología estándar para este tipo de proceso llamada CRISP-DM. Las fases documentadas son: conocimiento del negocio, conocimiento de los datos, preparación de los datos y finalmente el modelado.

OBJETIVOS

General

Documentar un proceso de minería de datos que sirva de guía para futuros profesionales de la ingeniería de la información, en el descubrimiento de conocimiento que se utiliza en la toma de decisiones con respecto a la gestión del talento humano.

Específicos

1. Demostrar la efectividad de la utilización de tecnologías de la información en torno a investigaciones de ciencias humanas, como lo es el tema de gestión de talento humano dentro de las organizaciones.

2. Realizar una revisión bibliográfica sobre el proceso de extracción de conocimiento en las bases de datos y sus diferentes técnicas para la búsqueda de información.

3. Diseñar, documentar e implementar un proceso de minería de datos sobre un conjunto de mediciones de clima organizacional, con el objetivo de descubrir información de utilidad para las empresas interesadas en la gestión del recurso humano.

INTRODUCCIÓN

El talento humano se ha convertido en una de las principales ventajas competitivas en los negocios. De ahí deriva la importancia de conocer las expectativas, motivaciones, necesidades y niveles de satisfacción de los colaboradores de una empresa para alcanzar una máxima productividad a nivel organizacional.

La evaluación de clima organizacional es un método ampliamente utilizado por las empresas, para evaluar la percepción de los colaboradores con respecto a los factores que inciden en la satisfacción, por ejemplo: la comunicación, el liderazgo, motivación, identificación con la empresa, etc. Generalmente esta información es analizada y presentada por medio de reportes e indicadores, los cuales se obtienen de un análisis estadístico descriptivo de los datos de las encuestas. Debido al gran volumen de datos y al número de variables involucradas que se manejan en estas evaluaciones, hay información implícita de gran valor que no se analiza, en la mayor parte de casos debido a los altos costos que implicaría realizarlo por medio de análisis estadísticos tradicionales.

La minería de datos es un conjunto de técnicas eficientes de análisis de información, que se utiliza para extraer conocimiento de grandes volúmenes de datos que puede ser útil para la empresa, institución o investigador que la realiza. Utiliza técnicas como la creación de modelos matemáticos y estadísticos, ejecución de algoritmos supervisados y no supervisados y algoritmos de inteligencia artificial por mencionar algunos.

El presente trabajo de graduación tiene como propósito diseñar, documentar y ejecutar un proceso completo de minería de datos por medio del cual se encuentren patrones y modelos de información sobre clima organizacional, los cuales permitan a las empresas tener un mejor entendimiento acerca de las necesidades y factores que un colaborador percibe como elementos de un buen clima organizacional, en específico determinar el perfil de los colaboradores más contentos y los más descontentos. Lo cual ayudará a la administración a identificar acciones necesarias para aumentar el nivel de satisfacción y en consecuencia, el rendimiento de la organización.

El análisis de datos se realiza con la información de una base de datos proporcionada por una empresa consultora guatemalteca dedicada a realizar mediciones de clima organizacional. La base de datos inicial consta de 63 evaluaciones que incluyen aproximadamente 13,000 encuestas realizadas a colaboradores de distintas empresas guatemaltecas entre los años 2010 y 2012.

Al finalizar el análisis se presentan los modelos de datos encontrados, así como la respectiva interpretación de resultados para cada modelo.

1. MARCO TEÓRICO

1.1. Clima organizacional

En este capítulo se presentan los principales conceptos asociados a las evaluaciones de clima organizacional, tales como la definición, metodología y los instrumentos de medición.

1.1.1. Definición

En el campo de la psicología industrial se han desarrollado varios enfoques sobre el concepto de clima organizacional, el que más utilidad ha tenido es el que emplea como elemento fundamental las percepciones que el trabajador tiene de las estructuras y procesos que ocurren en un medio laboral.

A continuación se mencionan algunos aspectos que permiten aclarar el concepto de clima organizacional:

- El clima organizacional se refiere a las características del medio ambiente de trabajo, las cuales son percibidas directa o indirectamente por los trabajadores. Entre ellas se pueden mencionar: el trato del jefe hacia sus subordinados, la relación entre los colaboradores de la empresa, proveedores o clientes.

- El clima organizacional tiene influencia en el comportamiento laboral.

- Las características de la organización que componen el clima cambian muy poco en el tiempo e identifican y distinguen una organización de otra.

Goncalves lo define de la siguiente manera: “El clima organizacional es un fenómeno que media entre los factores del sistema organizacional y las tendencias motivacionales que se traducen en un comportamiento que tiene consecuencias sobre la organización (productividad, satisfacción, rotación, etc.).”[1]

Cuando una empresa tiene un clima organizacional positivo se observan los siguientes comportamientos: empoderamiento, compromiso, productividad, baja rotación, satisfacción, adaptación, innovación, entre otras.

Por el contrario, las consecuencias de un clima organizacional negativo son: baja productividad, inadaptación, alta rotación, falta de compromiso, ausentismo, carencia de innovación.

1.1.2. Importancia del clima organizacional

Actualmente la gestión del talento humano y la calidad del mismo, se han convertido en unas de las principales ventajas competitivas en los negocios. Es importante conocer las expectativas, motivaciones, necesidades y niveles de satisfacción de los colaboradores de una empresa para alcanzar una máxima productividad a nivel organizacional.

[1] GONCALVES, Alexis P. *El clima como término organizacional.* [en línea] <http://moodle.unid.edu.mx/> [Consulta: 09 de mayo de 2013].

Contar con personal satisfecho y motivado es determinante para el éxito de las organizaciones, por lo que conocer en forma clara y detallada los niveles de satisfacción que los colaboradores tienen en los distintos ámbitos de su actividad es una tarea de suma importancia para gestionar planes y proyectos de mejora.

De acuerdo con una investigación realizada en el 2012 por la firma Great Place To Work ® se logró demostrar el valor y la importancia de tener un ambiente y clima organizacional positivo. El estudio determinó cuales son las 25 empresas con mejor clima organizacional a nivel mundial.[2] A continuación se enumeran las conclusiones más relevantes del estudio:

- En lo referente a lo económico, en este estudio se descubrió que el promedio de ingresos de las empresas que estaban en la lista de las empresas con mejor clima organizacional se elevó en un índice del 9% anual.

- Una empresa con un buen clima organizacional atrae y permite retener el mejor talento humano dentro de sus filas. Estas 25 empresas generaron 120,000 empleos a nivel mundial. Y el índice de rotación de personal de las 25 mejores empresas es del 8% anual, comparado con un 9.1% anual que es el índice promedio en la industria estadounidense.

- Con respecto a la cultura de estas 25 empresas, el empoderar a sus colaboradores y apoyarlos tanto en el ámbito personal como profesional, resulta en un catalizador de la creatividad e innovación dentro de los procesos y productos que ofrecen, lo cual les permite crecer y prosperar.

[2] Great Place To Work. *Tendencias del 2012*. [en línea] <http://www.greatplacetowork.com/best-companies/worlds-best-multinationals/2012-trends> [Consulta: 10 de junio de 2013].

Específicamente conocer el clima organizacional de una empresa nos permite:

- Medir la satisfacción de los colaboradores en el trabajo.

- Identificar los motivos por los que los colaboradores se sienten motivados, o no, al desempeñar sus actividades laborales.

- Obtener retroalimentación de los procesos que determinan los comportamientos organizacionales.

- Introducir cambios planificados en las actitudes o conductas de los colaboradores.

1.1.3. Evaluación de clima organizacional

La evaluación del clima organizacional, se fundamenta en la medición de la percepción de los colaboradores sobre los aspectos que influyen en su motivación y por ende en la productividad de la organización. Utilizando técnicas adecuadamente diseñadas, se permite a los colaboradores expresar su opinión de cómo funciona la organización en sus distintos aspectos y cómo se sienten en ella.

Existen varios instrumentos de medición que se pueden utilizar[3]:

- Grupos de enfoque**:** son reuniones con grupos de colaboradores dirigidas por un experto quien plantea temas asociados al clima

[3] SIBAJA, Nelson. *Instrumentos para medir el clima organizacional.* [en línea] <http://www.slideboom.com/presentations/220161/Instrumentos-para-medir-el-clima-organizacional> [Consulta: 19 de mayo de 2013].

organizacional con el objetivo de recabar la opinión de los empleados acerca de estos temas.

- Buzones de sugerencias: es una alternativa para recopilar información de forma anónima que busca recopilar los puntos de vista de los empleados. Tiene la desventaja de ser poco estructurada.

- Encuesta: debe ser planteada de acuerdo los factores de clima que se quieren identificar, es aplicada de forma anónima a una muestra significativa de la población dentro de la empresa.

- Análisis documental: a partir de análisis de indicadores de recursos humanos como el nivel de ausentismo de la empresa, porcentaje de rotación de colaboradores, quejas recibidas sobre salarios.

En este trabajo de investigación se centrará en analizar los resultados de mediciones que utilizan la encuesta como instrumento de medición.

1.1.3.1. Metodología para la evaluación de clima organizacional

La metodología generalmente utilizada consta de las siguientes fases[4]:

- Definir estructura organizacional: consiste en identificar las áreas de trabajo en las cuales serán segmentados los colaboradores, con el objetivo de obtener resultados específicos por área que permitan tomar

[4] CALLEJA PALMA, Carmen. *La Evaluación del clima organizacional como alternativa de mejora de la productividad laboral.* [en línea] <http://www.bibliotecadigital.uson.mx/bdg_tesisIndice.aspx?tesis=21317> [Consulta: 4 de abril de 2013].

acciones concretas para la mejora del clima organizacional. Es importante establecer la logística de este proceso, definiendo las fechas y los grupos que se conformarán para la aplicación de las encuestas.

- Diseño del instrumento de medición (encuesta): en esta fase se definen los factores, también llamados variables, de clima que se evaluarán. Una vez establecidos los factores se procede a redactar las preguntas o enunciados con los cuales se medirá cada factor. Se debe realizar una revisión y adaptación del instrumento que se adecúe a las necesidades y características de la empresa.

- Comunicación y sensibilización: debe establecerse una estrategia de comunicación con el fin de informar y sensibilizar a los colaboradores acerca de la evaluación de clima organizacional. Debe comunicarse cuales son los objetivos del estudio, cómo y cuándo se realizará y recalcar la confidencialidad con la que se manejará la información con el fin de garantizar el anonimato de las encuestas.

- Aplicación de las encuestas: de acuerdo al tipo de empresa, será factible que un porcentaje de los colaboradores realice la encuesta vía electrónica, en la cual se habilita un *link* para que el encuestado acceda por medio de internet y proceda a completar la encuesta. Con el resto del personal se aplican encuestas en hojas de papel y posteriormente se realiza la digitación de las mismas en un medio electrónico para su posterior análisis.

- Análisis de resultados: una vez recolectada toda la información en un medio electrónico, se procede a realizar el análisis de los resultados. Este análisis generalmente incluye el cálculo del nivel de satisfacción a

nivel de empresa y por área, así como identificar los factores mejor y peor calificados por los colaboradores. Deben elaborarse conclusiones y recomendaciones generales.

- Presentación de resultados: es importante comunicar los resultados de la evaluación de clima organizacional. En general se comparte el resultado global de satisfacción, mientras que a los gerentes, jefes y supervisores deberá entregarle los resultados del área que lideran para que puedan elaborar planes de acción concretos de acuerdo a los resultados obtenidos para mejorar el clima organizacional.

Uno de los aspectos más importantes que se debe considerar en la metodología para la aplicación de una encuesta de clima organizacional es el anonimato de los colaboradores encuestados. Esto asegura que los colaboradores contesten con una mayor sinceridad a las preguntas o afirmaciones planteadas.

1.1.4. Instrumento de medición: encuesta

Una encuesta es: "un estudio en el cual el investigador obtiene los datos a partir de realizar un conjunto de preguntas normalizadas dirigidas a una muestra representativa o al conjunto total de la población estadística en estudio, formada a menudo por personas, empresas o entes institucionales, con el fin de conocer estados de opinión, características o hechos específicos."[5]

La encuesta utilizada para medir clima organizacional permite cuantificar la percepción de los colaboradores con respecto a los factores o variables de

[5] *¿Qué es una encuesta?* [en línea] <http://www.portaldeencuestas.com/que-es-una-encuesta.php> [Consulta: 8 de junio 2013].

clima organizacional que hayan sido planteadas. En este tipo de encuestas se utilizan preguntas de respuesta cerrada, en la cual los encuestados deben elegir una de las opciones que se les plantean.

Las encuestas deben contener los siguientes elementos:

- Factores: o variables del clima organizacional que serán evaluadas. En la tabla I se describen los factores que generalmente forman parte de una evaluación de este tipo.

Tabla I. **Factores de una evaluación de clima organizacional**

Factor	Aspectos evaluados
Ambiente de Trabajo y Motivación	Evalúa condiciones emocionales del colaborador para un desempeño y productividad adecuados, así como la satisfacción y motivación individual y el reconocimiento por el trabajo bien hecho.
Capacitación y Desarrollo	Evalúa la opinión del personal en relación con los programas de capacitación y desarrollo, así como la motivación para aprovechar las oportunidades de formación.
Comunicación	Evalúa varios aspectos de la comunicación en la organización: Relación jefe-subalterno, comunicación entre compañeros del área de trabajo, con otras áreas de la Empresa, claridad de objetivos y cambios.
Identificación e imagen Empresarial	Evalúa el sentido de pertenencia, así como la opinión sobre estabilidad, imagen de la Empresa, confianza en que se cumplen los compromisos y grado en que se siente el apoyo de la misma hacia sus colaboradores.
Liderazgo	Evalúa el estilo de liderazgo: Trato que los jefes dan a sus colaboradores, inclusión, respeto, orientación, reconocimiento y participación.
Organización y Cambio	Evalúa la opinión sobre los procesos de cambio, forma de organización, definición de roles y distribución de cargas de trabajo.
Seguridad, Orden y Limpieza	Evalúa la opinión del personal en relación con las condiciones de trabajo, aspectos de seguridad, orden, limpieza, ambientes físicos.
Trabajo en Equipo	Evalúa la calidad de la relación entre áreas, coordinación, cooperación y apoyo para logro de objetivos, la integración con los compañeros de un mismo departamento y trabajo en equipo.

Fuente: elaboración propia, basada en los factores propuestos por la empresa Estratek.

- Reactivos: son afirmaciones o enunciados que plantean situaciones o comportamientos, que pueden ser observados por el colaborador y para los cuales debe indicar si está de acuerdo o en desacuerdo. Cada reactivo está asociado a uno de los factores que evalúan. En la redacción de los reactivos debe evitarse que contengan ambigüedad; deben expresar aprobación o rechazo con respecto a situaciones que representen las dimensiones que se están midiendo.

- Escala de medición: una escala es "un conjunto de reactivos verbales ante los cuales un individuo responde expresando grados de acuerdo o desacuerdo, o algún otro modo de respuesta. Los reactivos de escala tienen alternativas fijas y colocan sobre algún punto de la escala al individuo que responde"[6]. Comúnmente en los estudios de ciencias sociales y psicológicas se utiliza la escala tipo Likert, en la cual la respuesta especifica un nivel de desacuerdo o acuerdo con una declaración o reactivo. Un ejemplo se muestra en la tabla II.

Tabla II. **Ejemplo de escala tipo Likert**

Tipo de Reactivo	Nunca	A veces	Casi siempre	Siempre	No Aplica
Reactivos Positivos	1	2	3	4	N/A
Reactivos Negativos	4	3	2	1	N/A

N/A indica que la elección de este ítem no tiene ninguna incidencia en los resultados

Fuente: elaboración propia.

- Datos demográficos: con el objetivo obtener información relevante sobre la percepción del clima organizacional en distintos grupos demográficos

[6] *Métodos de encuesta: entrevistas y cuestionarios.* [en línea] <http://www2.udec.cl/~gnavarro/2001_1/ienc.html> [Consulta: 10 de mayo 2013].

de la empresa, se puede agregar un apartado donde se solicite información demográfica, siempre y cuando no comprometa la confidencialidad del encuestado. Algunos ejemplos de datos demográficos utilizados en las encuestas de clima son: edad, sexo, tiempo de permanencia en la empresa, grado académico.

- Preguntas abiertas: es importante incluir una pregunta abierta que permita recolectar comentarios de temas que los colaboradores consideren importantes para mantener o mejorar el clima organizacional de su área o empresa. Este tipo de preguntas brinda información cualitativa acerca del clima de la empresa.

1.1.5. Situación actual de las herramientas para medir clima organizacional y sus limitantes

Toda la información recabada en la medición de clima organizacional, es analizada con base en la perspectiva del consultor o a requerimiento de los clientes. La mayoría de las herramientas tienen definido una serie de reportes e indicadores, los cuales se obtienen a partir de un análisis estadístico descriptivo de las encuestas.

Las tecnologías de información que se utilizan actualmente para analizar y presentar los resultados de una medición de clima organizacional son las siguientes:

- Hojas electrónicas de cálculo.

- Software genérico para aplicar encuestas.

- Módulo de recursos humanos del *ERP* (sistema de planificación de recursos empresariales).

- Software especializado para la medición de clima organizacional.

En las tecnologías mencionadas, el análisis de resultados tiene un enfoque estadístico sencillo, valga decir, que en la mayoría de los casos este permite responder a los planteamientos básicos de un estudio de clima organizacional.

De acuerdo a una revisión realizada por el autor en conjunto con un consultor de la empresa Estratek, se identificaron cuáles son los resultados e indicadores que comúnmente se obtienen con las herramientas enumeradas anteriormente:

- Índice de satisfacción de los colaboradores, tanto a nivel global de la empresa como segmentado por área de trabajo. Este índice se calcula con una media aritmética de las respuestas obtenidas en las encuestas aplicadas.

- Índice de satisfacción de cada uno de los factores o variables de clima organizacional que evaluaron en el instrumento de medición. Este índice se calcula con una media aritmética de las respuestas asociadas a cada factor de las encuestas aplicadas.

- Identificar los factores del clima organizacional que representen una fortaleza y cuales necesitan ser mejorados. Se calcula el índice de satisfacción de los factores, se listan en orden descendente y se presenta el listado de los 3 índices más altos y los 3 índices más bajos.

- Comparar el índice de satisfacción entre las distintas áreas de trabajo de la empresa.

- Analizar los cambios en el clima organizacional de la empresa que han habido a través del tiempo. Se realiza un comparativo entre los estudios que se han realizado en anteriores ocasiones para medir el impacto que tuvieron las acciones de mejora planteadas entre una evaluación y otra.

Debido al gran volumen de datos y el número de variables involucradas que se encuentran en un estudio de clima organizacional, hay información importante que es recolectada y que en las herramientas anteriormente descritas, no se analiza:

- Correlación entre datos demográficos y el índice de satisfacción de empleador.

- Perfil demográfico de los colaboradores con mejor índice de satisfacción.

- Independencia o dependencia de los factores analizados en el clima organizacional.

1.2. Minería de datos

A continuación se presenta las bases teóricas involucradas en los procesos de minería de datos, las cuales se utilizan como parte de los sistemas de información para la toma de decisiones.

1.2.1. Definición

La minería de datos es un proceso mediante el cual se extrae conocimiento implícito, previamente desconocido y potencialmente útil, a partir de grandes volúmenes de datos; empleando técnicas de inteligencia artificial, análisis matemático y estadístico para descubrir patrones y tendencias en los datos. [7]

Se les denomina patrones a las relaciones que existen entre los elementos de los datos analizados. Los patrones son de interés, si son confiables, novedosos y útiles respecto al conocimiento que generan y el acoplamiento con los objetivos del análisis.

La aplicación de la minería de datos se puede observar en los siguientes escenarios:

- Riesgo y probabilidad: determinación de puntos de equilibrio para escenarios de riesgo, asignación de probabilidades a diagnósticos.

- Segmentación: consiste en crear grupos a partir de un conjunto de datos basándose en la similitud de sus atributos.

- Búsqueda de secuencias y recomendaciones: análisis de secuencias y patrones de compras de un consumidor en el supermercado y recomendaciones de compra.

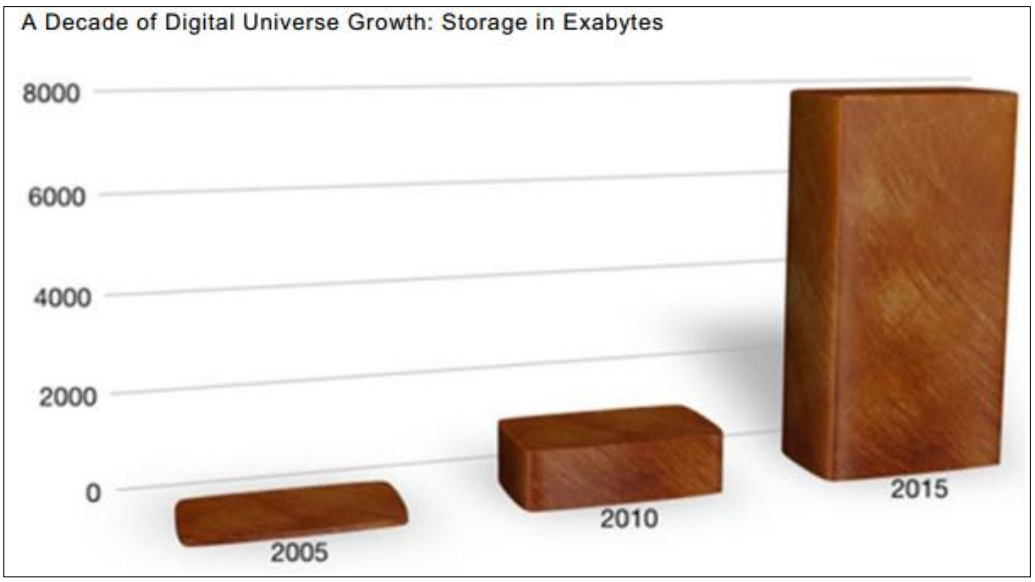
- Pronóstico: estimación de ventas, predicción de índices bursátiles.

[7] GARCIA SANTIESTEBAN, David. *Minería de datos para la detección de patrones criminalísticos en Guatemala.* Universidad de San Carlos de Guatemala, 2012.

1.2.2. Ventajas de utilizar minería de datos

La minería de datos ha tenido un gran auge en la era de la información debido a que la cantidad de datos que hay disponibles cada año crece exponencialmente. De acuerdo a un estudio patrocinado por la empresa multinacional EMC (dedicada al servicio de almacenamiento de datos y seguridad de la información entre otros) asegura que la cantidad de la información digital crecerá 44 veces del año 2009 al 2020.[8] En la figura 1 se muestra una tendencia del crecimiento de la información digital en *exabytes* para la década del 2005 al 2015.

Figura 1. **Proyección del crecimiento de información digital**



Fuente: Gantz, John. *Extracting value from chaos.* http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. Consulta: julio de 2013.

[8] *IDC proyecta un crecimiento de información del 45% para 2020*. [en línea] <http://cxo-community.com/articulos/estadisticas/82-tecnologias-infraestructura/3214-idc-proyecta-un-crecimiento-de-informacion-del-45-para-2020.html> [Consulta: 03 de junio de 2013].

Debido a este crecimiento de los datos, se hace es indispensable tener herramientas que permitan extraer conocimiento de la información disponible, que luego servirá de apoyo en los procesos de toma de decisiones que aporten al cumplimiento de los objetivos de las empresas.

A continuación se describen las principales ventajas de utilizar minería de datos como tecnología para el análisis de información:

- Permite descubrir información clave para el negocio, que normalmente está implícita y oculta dentro de las bases de datos con gran volumen de información, que de otra manera hubiera sido imposible encontrar o demasiado con un costo demasiado alto.

- Este tipo de proyectos se integra a la toma de decisiones estratégicas de la empresa.

- Ayuda a priorizar la toma de decisiones y acciones, mostrando los factores que tienen mayor incidencia en el alcance de un objetivo de negocio.

- Abre espacio para que se comparta información entre investigadores y los emprendedores de negocios.

- La minería de datos provee modelos de información descriptivos los cuales permiten la exploración automatizada de datos; ayuda a la comprensión de datos por medio de visualizaciones e identifican patrones, relaciones y dependencias que impactan en la rentabilidad del negocio (aumento de los ingresos, mejora de la productividad de los colaboradores, reducción de costos y gestión de riesgos).

- También genera modelos predictivos, los cuales permiten predecir escenarios futuros con base en relaciones no descubiertas e identificadas en la información actual. Por ejemplo, predecir los productos de la empresa pueden tener mayor rentabilidad basándose en los históricos de compra de los clientes.

1.2.3. Proceso de minería de datos

En la medida en que ha evolucionado e incrementado el uso de la minería de datos a nivel de instituciones y empresas, ha sido necesario definir una metodología que provea los lineamientos y mejores prácticas para llevar a cabo un proyecto de esta naturaleza.

En 1996, 3 empresas líderes de la industria (DaimlerBenz, SPSS y NCR) se unieron para realizar aportes en base a su experiencia en el tema de *Data Mining* y crearon un modelo de proceso llamado CRISP-DM (Cross Industry Standard Process for Data Mining), el cual describe una serie de fases y enfoques para abordar un proyecto de minería de datos.[9]

De acuerdo al estándar CRISP-DM, para la planificación del proyecto se debe tomar en cuenta la asignación porcentual de tiempo a cada fase del proyecto, que se muestra en la tabla III.

[9] IBM Corporation. *IBM SPSS modeler CRISP-DM guide*. [en línea] <ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf> [Consulta: 19 de junio de 2013].

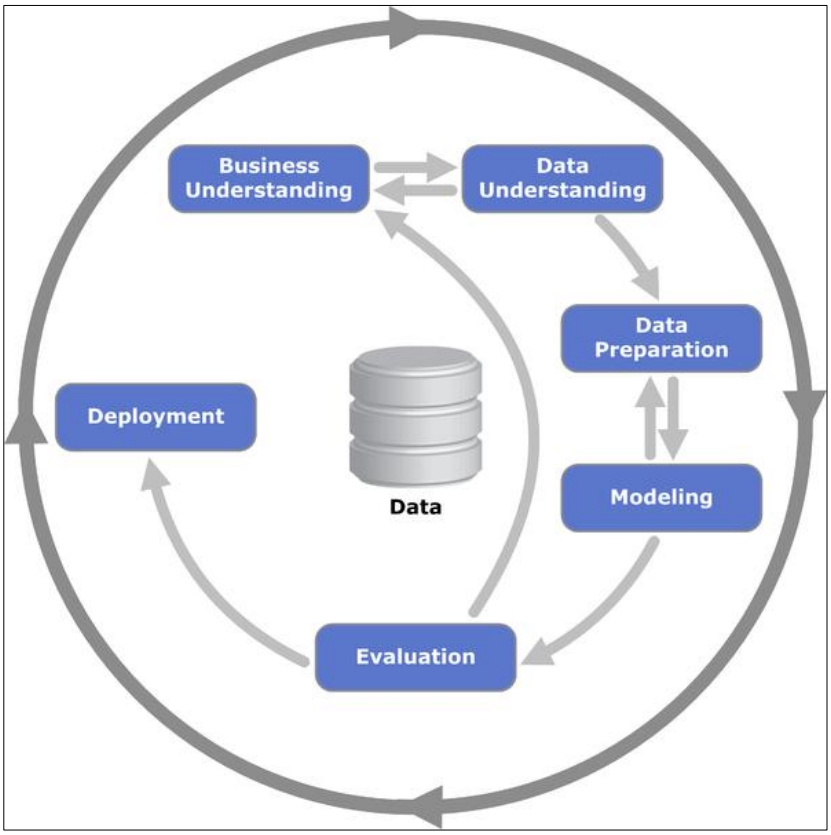
Tabla III. **Planificación proyecto de minería de datos**

% Tiempo Asignado	Fase
50 - 70 %	Preparación de los datos
20 - 30 %	Conocimiento de los datos
10 - 20 %	Modelado, evaluación y conocimiento del negocio
5 - 10 %	Despliegue del proyecto

Fuente: elaboración propia, basado en la metodología CRISP-DM.

En la figura 2 se puede observar el ciclo de vida de un proyecto de minería de datos que sigue la metodología CRISP-DM, este diagrama ayuda a entender las fases del proceso y provee una estrategia para ejecutar el proyecto. Las flechas entre las distintas fases indican las dependencias más importantes entre ellas. El círculo exterior enfatiza que es un proceso cíclico que se beneficiará de las experiencias aprendidas durante iteraciones anteriores.

Figura 2. **Diagrama del proceso de un proyecto de minería de datos**



Fuente: http://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png. Consulta: junio de 2013.

1.2.3.1. Fases del proceso CRISP-DM

- Conocimiento del negocio: en este paso se debe analizar los requisitos, analizar el contexto del problema, entender los objetivos del negocio desde una perspectiva no técnica y generar el plan del proyecto.

- Conocimiento de los datos: el objetivo de esta fase es familiarizarse con los datos tomando en cuenta los objetivos del negocio. Se realizan las

siguientes tareas con respecto a los datos: recopilación inicial de datos, descripción, exploración y verificación de la calidad de los mismos.

- Preparación de los datos: se deben realizar las tareas necesarias para construir el conjunto final de datos que se utilizarán con las herramientas de modelado. Se seleccionan los datos, se identifica y corrige información faltante, se realiza depuración de aquellos datos que tengan atributos incorrectos. También se puede realizar el enriquecimiento de datos, que consiste en agregar atributos a los datos ya existentes, con el propósito de satisfacer los requisitos del proyecto de minería de datos. Las técnicas de limpieza, transformación y reducción del número de dimensiones de los datos permiten asegurar la calidad de los mismos.

- Modelado: en este momento se seleccionan y aplican las técnicas de minería de datos, ajustando los parámetros a sus valores óptimos. Se generan pruebas, crean los modelos y se interpretan los resultados.

- Evaluación: a partir de los modelos generados en la fase anterior, se debe evaluar si estos son útiles y cumplen con los objetivos del negocio establecidos. Se realiza la evaluación de los resultados, revisión del proceso y se establecen las siguientes acciones a realizar.

- Despliegue: una vez que se han validado los modelos, es crucial explotar la utilidad de los mismos, integrándolos a las tareas de toma de decisiones del negocio. En esta fase se planifica el despliegue del proyecto de minería de datos, el monitoreo y mantenimiento del mismo. Se genera un informe final y se realiza una revisión del proyecto.

1.2.4. Modelos de minería de datos

De acuerdo al propósito a los modelos de minería de datos, estos pueden clasificarse en modelos descriptivos y modelos predictivos.

En la tabla 2 se observa una clasificación de técnicas de minería de datos de acuerdo al modelo que utilizan:

Tabla IV. **Clasificación de técnicas de minería de datos**

Tipos de Algoritmo	Modelo Utilizado
Clustering o segmentación de datos	Descriptivo
Estimación o regresión	Predictivo
Asociación o secuencia de patrones	Descriptivo
Clasificación	Predictivo

Fuente: elaboración propia, basada en la presentación *Data Mining,* Sangeeta Devadiga, 2007.

1.2.4.1. Modelos descriptivos

Los modelos descriptivos intentan encontrar patrones entre las relaciones de los datos y sus atributos. Estos modelos sintetizan todos los datos con el fin de proporcionar información con respecto a tendencias, segmentos y grupos que están presentes en la información buscada.

Estos modelos son utilizados muy a menudo en estudios de publicidad, ya que permite segmentar extensos grupos de consumidores de acuerdo a sus características homogéneas.

1.2.4.2. Modelos predictivos

Este tipo de modelos se utilizan para predecir respuestas futuras basándose en un análisis de datos históricos. Se enfocan en la construcción de un modelo de datos basado en la información existente. Cuando se tiene el modelo se utiliza como base para predecir otra variable que es relevante a los datos analizados.

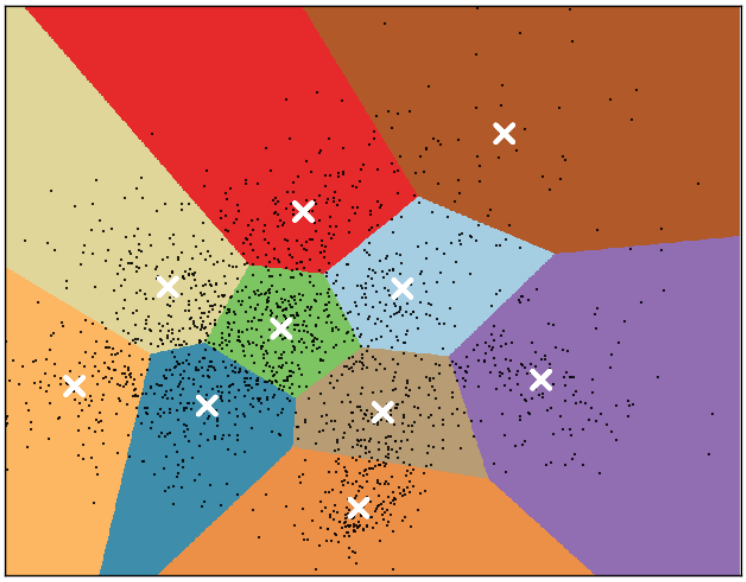
Los modelos predictivos son ampliamente utilizados por los mercadólogos para conocer los productos que están buscando los clientes, basándose en las tendencias actuales de compra, se pueden realizar predicciones de cuales productos nuevos pueden ser populares en el futuro.

1.2.5. *Clustering* o segmentación de datos

Este tipo de algoritmos tienen como objetivo agrupar conjunto de datos basándose en la similitud de sus atributos. Este agrupamiento es posible efectuarlo al identificar regiones de datos densamente pobladas, basándose en una medida de distancia que se establece al inicio de las iteraciones del algoritmo. A medida que se itera, se busca tener mayor similitud entre los elementos de cada grupo y menor similitud entre grupos.

En la figura 3 se puede ver una representación gráfica de los grupos creados a partir de la aplicación del algoritmo K-medias.

Figura 3. **Ejemplo de grupos creados a partir del algoritmo *k-means***



Fuente: http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html. Consulta: junio de 2013.

En la tabla V se presenta un resumen de los algoritmos más utilizados para realizar *clustering*.

Tabla V. **Algoritmos más utilizados en c*lustering***

Algoritmo	Objetivo
K-medias	Algoritmo de aprendizaje cuyo objetivo es encontrar la mejor división de n entidades en k grupos, para que la distancia total entre los elementos del grupo y su correspondiente centroide, sea minimizada.
Propagación de Afinidad	Un algoritmo que identifica ejemplares entre los puntos de datos y agrupaciones de los puntos de datos que se encuentran alrededor de estos ejemplares. Considera simultáneamente todos los puntos de datos como posibles, e intercambia mensajes entre los puntos de datos hasta que surja un buen conjunto de ejemplares y grupos.

Continuación tabla V.

Maximización de la Esperanza (EM)	Algoritmo de *clustering* en donde los puntos a clasificar se asumen como si estos fueran derivados de una mezcla de componentes *gaussianas* o *poisson* y el interés es encontrar los parámetros de dichas distribuciones.
DBSCAN	El algoritmo considera los grupos como áreas de alta densidad separadas por zonas de baja densidad. Al utilizar este punto de vista muy genérico los *clusters* encontrados pueden tener cualquier forma, al contrario de k-medias que asume que los grupos son de forma convexa.
Agrupación Jerárquica	Busca construir una jerarquía de agrupaciones construyendo grupos anidados y fusionándolas sucesivamente. Esta jerarquía de grupos se representa por un árbol donde su raíz es el único grupo que reúne todas las muestras y las hojas son grupos con una sola muestra.

Fuente: elaboración propia.

1.2.6. Estimación o regresión

Esta técnica es utilizada para generar un modelo que pueda predecir el valor de una clase, basado en los valores de los atributos de los datos. Por ejemplo: estimar número de hijos en un grupo familiar, determinar la probabilidad de que una transacción haya sido fraudulenta.

En la tabla VI se presenta un resumen de los algoritmos más utilizados para realizar proyectos que requieran estimación o regresión de los datos.

Tabla VI. **Algoritmos más utilizados para estimación o regresión**

Algoritmo	Objetivo
Árboles de Clasificación y regresión (CART)	Procedimiento recursivo de partición capaz de procesar atributos continuos y continuos como objetivos o predictores.

Continuación tabla VI.

Regresión lineal	Ajusta un modelo lineal para el conjunto de datos mediante el ajuste de un conjunto de parámetros con el fin de hacer que la suma de los residuos al cuadrado del modelo sea tan pequeña como sea posible.
Redes neuronales	Técnica de regresión no lineal, que es utilizada para reconocimiento de patrones y clasificación. Constituyen una técnica de modelización multivariada, es decir, pueden hacer predicciones de dos o más variables objetivo simultáneamente y en interacción, como también en cascada.

Fuente: elaboración propia.

1.2.7. Asociación

Esta técnica se utiliza para descubrir hechos que ocurren en común dentro del conjunto de datos que se está analizando. Es utilizado en decisiones de *marketing* para el análisis de la carretilla de compras, detección de intrusos entre otros.

Para esta técnica se utiliza reglas de asociación, las cuales son reglas que implican ciertas relaciones de asociación entre un conjunto de objetos en un conjunto de datos.

En la tabla VII se condensan los algoritmos más utilizados para construir modelos de asociación o frecuencia de patrones.

Tabla VII. **Algoritmos utilizados en asociación o frecuencia de patrones**

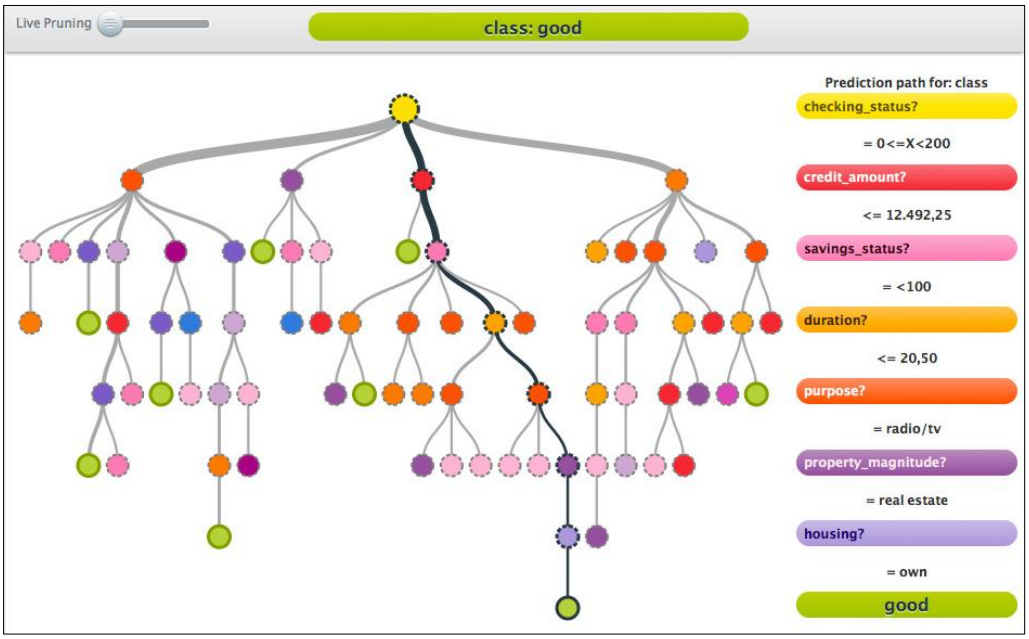
Algoritmo	Objetivo
A priori	Algoritmo de aprendizaje de reglas de asociación, diseñado para operar en bases de datos que contengan transacciones. Dado un conjunto de elementos, el algoritmo intenta encontrar los subconjuntos que tienen en común por lo menos un número X de elementos.
FP-Growth	Es un método eficiente y escalable para encontrar conjuntos de patrones frecuentes. Utiliza una estructura extendida de "árboles prefijo" para almacenar información crucial y comprimida de los patrones frecuentes llamados árbol de patrones frecuentes.
Eclat	Este algoritmo utiliza intersecciones para calcular el apoyo de un conjunto de elementos candidatos, evitando la generación de subconjuntos que no existen en el árbol prefijo.

Fuente: elaboración propia.

1.2.8. Clasificación

Esta técnica consiste en utilizar los atributos de un caso para clasificarlo en una clase definida previamente. Para la construcción de este modelo se tiene un conjunto de datos de entrenamiento, donde cada dato tiene un conjunto de atributos y está asignado a una clase específica. Mediante un método supervisado se analiza el conjunto de entrenamiento y se propone un modelo para cada clase utilizando los atributos de los datos. A partir de los modelos generados para cada clase se pueden clasificar otros conjuntos de datos.

Figura 4. **Árbol de decisión para clasificar clientes que aplican crédito**



Fuente: http://decisiontrees.net/wp-content/uploads/2012/05/bigml.jpg. Consulta: junio de 2013.

Uno de los tipos de algoritmo más utilizado para realizar clasificación son los árboles de decisión (ver figura 4). El objetivo es crear un modelo que prediga el valor de una variable a partir del aprendizaje de simples reglas de decisión que son inferidas de las características de los datos.

En la tabla VIII se describen los algoritmos más utilizados para construir modelos de clasificación de datos.

Tabla VIII. **Algoritmos utilizados en clasificación**

Algoritmo	Objetivo
Naïve Bayes	Este algoritmo está basado en la aplicación de teorema de Bayes (a partir de la estadística bayesiana) con las hipótesis de independencia fuertes (Naive). Dado un conjunto de objetos, que pertenecen a una clase conocida, construye una regla que permite asignar objetos futuros a una clase. Es especialmente adecuado cuando la dimensionalidad de las entradas es alta.
Árboles de Decisión	El objetivo es crear un modelo que prediga el valor de una variable a partir del aprendizaje de simples reglas de decisión que son inferidas de las características de los datos.
Análisis Linear de Discriminantes	Es una técnica estadística multivariante cuya finalidad es describir (si existen) las diferencias entre g grupos de objetos sobre los que se observan p variables (variables discriminantes).
Máquinas de Vectores (SVM)	Mediante el aprendizaje, este algoritmo trata de encontrar la mejor función de clasificación para distinguir en miembros de distintas clases.
AdaBoost	Emplea métodos que utilizan múltiples *learners* para resolver un problema.
kNN	Este algoritmo memoriza el conjunto de datos de entrenamiento y ejecuta una clasificación sólo si los atributos del objeto de prueba coinciden exactamente con los ejemplos del entrenamiento.

Fuente: elaboración propia.

1.3. Estratek y las evaluaciones de clima organizacional

Estratek es una empresa guatemalteca fundada en el 2010, la cual es gestionada por profesionales con amplia experiencia en el área de recursos humanos y de tecnologías de la información.

Estratek se dedica a proveer herramientas de análisis para la toma de decisiones estratégicas en gestión del talento y la salud emocional de la organización, como lo son evaluaciones de clima organizacional y evaluaciones del desempeño.

La herramienta de clima organizacional que posee la empresa está basada en la aplicación de encuestas que son aplicadas de manera física y electrónica. La información física es posteriormente digitalizada en un repositorio de datos virtual. A partir de este repositorio se provee a los consultores una plataforma web de reportes que permite analizar los resultados de la evaluación de clima organizacional de la empresa, de cada área o departamento evaluado, los factores y reactivos utilizados en la encuesta.

Está empresa proveerá la información de 63 estudios de clima organizacional para que en el presente estudio se realice, la minería de datos de acuerdo a los objetivos y lineamientos que se plantearán en el segundo capítulo de este trabajo.

2. PROCESO DE MINERÍA DE DATOS APLICADO A CLIMA ORGANIZACIONAL

2.1. Metodología de trabajo

Para el desarrollo del presente trabajo de investigación se utilizará la metodología para procesos de minería de datos CRISP-DM (de la cual se realizó una revisión bibliográfica en el capítulo 1). De acuerdo a la metodología, la inclusión de fases y tareas específicas se determina de acuerdo a las necesidades específicas del proyecto. Para el caso puntual de esta investigación, no se realizará la fase de despliegue debido a que no es necesario implementar un el plan de despliegue, monitoreo y mantenimiento para lograr cumplir los objetivos planteados.

2.2. *R* como herramienta de software para la minería de datos

La herramienta que se seleccionó para realizar el análisis de datos se llama *R*, un software estadístico creado por Ross Ihaka y Robert Gentleman de la Universidad de Aukland en Nueva Zelandia y está diseñado para el análisis de datos, gráficos y análisis estadísticos.

La elección de la plataforma computacional *R* para el desarrollo de esta tesis se debe a que es una plataforma bastante conocida, tiene una facilidad para programar, tiene un buen rendimiento, extensa documentación y su uso está siendo ampliado a campos como bioinformática, finanzas entre otros.

A continuación se detallan otras ventajas adicionales y también desventajas que posee el programa *R*.

2.2.1. Ventajas

- Software de código abierto y multiplataforma (Windows, Linux y MacOS).

- Completamente programable y extensible por medio de instalación de paquetes que proveen flexibilidad en el análisis de datos.

- Existe amplia documentación para la utilización de *R* para minería de datos.

- Tiene una comunidad de desarrollo activa, que actualiza constantemente el software.

- Existen actualmente 6 paquetes especializados que incluyen alrededor de 40 algoritmos implementados para el software *R*, que permiten desarrollar técnicas de *data mining* como*:*
  - Reducción de dimensionalidad
  - Clasificación
  - *Clustering* o segmentación de datos
  - Asociación

2.2.2. Desventajas

- Debido a que la interacción con el usuario se realiza por medio de una interfaz de comandos y no una interfaz gráfica, es necesario conocer o

tener un documento de referencia de los comandos que se desean utilizar.

- Requiere invertir un considerable tiempo inicial para obtener resultados observables.

2.3. Conocimiento del negocio

El primer paso para iniciar el proyecto, es explorar lo que la organización espera obtener de la minería de datos. Conocer las razones a nivel de negocio, por las cuales se realiza la minería de datos ayudará a que todos los involucrados tengan la misma perspectiva antes de gastar valiosos recursos.

2.3.1. Determinando los objetivos del negocio

El proceso de definición de objetivos, permite descubrir al inicio del proyecto, factores importantes que influyen en el resultado final.

2.3.1.1. Antecedentes

Toda empresa necesita del recurso humano para poder llevar a cabo el cumplimiento de sus metas. El departamento de recursos humanos es el encargado de gestionar lo relativo a los colaboradores: reclutamiento, selección, contratación, capacitación y establecimiento de estrategias de gestión que faciliten alcanzar los objetivos de la organización y el incremento en la productividad de los colaboradores.[10]

[10] *Administración de Recursos Humanos.* [en línea] <http://www.gestiopolis.com/organizacion-talento/administracion-de-recursos-humanos-.htm> [Consulta: 10 de julio de 2013].

Contar con personal satisfecho y motivado es determinante para el éxito de las organizaciones, por lo que conocer en forma clara y detallada los niveles de satisfacción que los colaboradores tienen en los distintos ámbitos de su actividad, es una tarea de suma importancia para gestionar planes y proyectos de mejora. La herramienta que permite conocer ese nivel de satisfacción es la evaluación del clima organizacional.[11]

Las empresas dedicadas a la consultoría en clima organizacional, a través del tiempo han desarrollado modelos y metodologías basados en marcos conceptuales y su propia experiencia, lo cual les permite ofrecer una gestión eficiente en las mediciones de satisfacción de los empleados. Muchas de estas empresas consultoras poseen encuestas modelo, software especializado para la captura de datos y análisis de resultados y consultores de recursos humanos especializados en el tema de clima organizacional.

Como resultado de las mediciones de clima organizacional que estas empresas realizan, se generan bases de datos históricas de gran volumen que contienen la captura de todas las encuestas realizadas a los distintos clientes que contratan este tipo de servicios. Esos datos que en su momento fueron analizados desde la perspectiva de la empresa cliente, tienen un gran valor científico que está implícito y oculto y que es susceptible de ser descubierto si es analizado desde otras perspectivas y con otro tipo de herramientas. Por ejemplo al seleccionar datos de clima organizacional de empresas que pertenecen a una determinada región geográfica o a un sector empresarial específico, puede resultar valioso analizar lo siguiente:

- Las tendencias a nivel macro de clima organizacional.

[11] PÉREZ, Isabel y MALDONADO, Marisabel. *Clima organizacional y gerencia: inductores del cambio organizacional.* [en línea] <http://dialnet.unirioja.es/descarga/articulo/2310289.pdf> [consulta: 20 de junio de 2013].

- Intentar determinar si los niveles de satisfacción para cierto segmento son comparables, tienen un comportamiento de distribución normal o se ajustan a algún modelo de distribución de datos.

- Validar y demostrar con rigor estadístico algunos hechos que se observan empíricamente, como por ejemplo: la dependencia entre ciertos factores que se miden el clima (liderazgo, motivación, comunicación).

Todos estos beneficios no son por completo ajenos a los consultores que se dedican a este campo, sin embargo existen limitantes que no permiten realizar ese tipo de exploraciones, entre ellos se pueden mencionar:

- Poco conocimiento acerca de las tecnologías disponibles para analizar bases de datos de gran volumen.

- Consultores con mucho conocimiento conceptual acerca del tema de clima organizacional, pero con limitado conocimiento de análisis estadístico que involucre el desarrollo o utilización de algoritmos de minería de datos.

- A nivel general, las empresas consultoras han encontrado una zona de *confort* en la cual los clientes están satisfechos con la información que les proporcionan los análisis de resultados actuales. Donde esa satisfacción por parte de los clientes, no implica necesariamente que los resultados obtenidos estén generando planes de acción que impacten positivamente en el clima organizacional. Muchas veces la información brindada ayuda a “apagar fuegos”, más no a realizar cambios a nivel cultural lo que podría impactar en mejoras al clima a largo plazo.

2.3.1.2. Objetivos del negocio

El principal objetivo es establecer un proceso que permita aprovechar la información histórica de evaluaciones de clima organizacional recolectada por la empresa especializada Estratek, con el fin de extraer conocimiento que permita proveer nuevos modelos de acción que permitan incrementar el nivel de satisfacción de los trabajadores de las empresas.

Los objetivos específicos de este trabajo de investigación son:

- Identificar cuáles son los perfiles de los colaboradores que perciben mayor satisfacción en el desarrollo y ambiente laboral en que se desenvuelven, esto con el objetivo de identificar los candidatos ideales para el tipo de negocio y puestos que ofrecen.

- Determinar los factores o variables de clima organizacional que tienen dependencia entre ellas. Esto permitirá diseñar soluciones o acciones que incidan en mejorar la satisfacción en más de un factor, haciéndolas de esta manera soluciones más efectivas.

- Documentar un proceso de minería de datos que pueda servir de guía para futuros profesionales de la ingeniería de la información, en el descubrimiento de conocimiento que se utiliza en la toma de decisiones, con respecto a la gestión del recurso humano e introducción de cambios en la cultura organizacional a nivel empresarial.

2.3.1.3. Criterios de éxito

El presente trabajo de investigación podrá ser considerado exitoso si:

- Logra descubrir una segmentación válida de los grupos demográficos que tienen una mayor satisfacción en el trabajo que desempeñan.

- Demuestra la dependencia o independencia de algunos factores de la evaluación de clima organizacional, con respecto a las factores que tradicionalmente se les denomina principales o de mayor impacto en el como lo es el liderazgo.

- Documenta cada una de las fases, tareas y análisis ejecutados en el proceso de minería de datos aplicada a estudios de clima organizacional.

2.3.2. Evaluando el contexto

Ahora que se tienen claros los objetivos a nivel de negocio del proyecto, se realizará un análisis de los recursos, alcance y otros factores que deben considerarse al determinar las metas de análisis de datos y en el desarrollo del plan del proyecto.

2.3.2.1. Recursos

Recursos con los que se disponen para llevar a cabo el proyecto de minería de datos:

- Humanos: investigador, asesor de tesis, consultores expertos de la empresa Estratek.

- Tecnológicos: motor de base de datos con información recolectada, software *R* para desarrollar las técnicas de minería de datos y acceso a internet para consulta de documentación del software y descarga de librerías necesarias para el software *R*.

- Datos para análisis: Estratek pondrá a disposición una base de datos con 63 estudios de clima organizacional realizados en los últimos 3 años.

- Soporte conceptual y técnico: como parte del desarrollo de esta investigación se realizó un marco teórico que describe, explica y documenta los principales concepto y técnicas relacionadas con clima organizacional y minería de datos.

2.3.2.2. Requerimientos, supuestos y alcance

- Requerimientos: todas las herramientas informáticas que se utilizarán para el análisis de datos deben ser de código abierto. Esto con la finalidad que cualquier profesional interesado en el estudio, no tenga un impedimento inicial para acceder a los resultados y tenga la capacidad de emularlos con fines de ampliar los conocimientos generados con este tipo de análisis.

- Supuestos: los datos que serán proveídos por Estratek para el análisis son confiables y son susceptibles de sufrir modificaciones, con el fin de preparar la información y garantizar la calidad de los mismos en el proceso de minería de datos.

- Restricciones: la empresa Estratek y el investigador suscriben un acuerdo de confidencialidad, donde se prohíbe compartir los datos con

terceras personas, no se les puede dar otro uso adicional al de utilizarse para los objetivos antes descritos y se prohíbe publicar datos como: nombres de las empresas evaluadas, año en que se evaluaron. En el anexo de este trabajo se incluye una copia del contrato de confidencialidad firmado.

2.3.3. Determinando las metas de minería de datos

Después de realizar un análisis y entendimiento de los objetivos del negocio, a continuación se plantean los objetivos en términos de minería de datos.

2.3.3.1. Objetivos de minería de datos

Al concluir el presente estudio se debe cumplir las siguientes metas:

- Desarrollar un modelo de segmentación o *clustering*, que permita segmentar la información de registros demográficos de las evaluaciones, para identificar los grupos de colaboradores con mayor satisfacción y sus características demográficas. Es deseable encontrar conclusiones como las siguientes:

    - Qué relación tiene el puesto (autonomía de decisiones) con la satisfacción del empleado.
    - Si el tiempo de permanencia dentro de la empresa o la edad influyen en la satisfacción laboral que goza el colaborador.
    - Las características demográficas que no tienen ningún impacto en el clima organizacional.

- Identificar si existe dependencia entre los factores que determinan el clima organizacional de una empresa. Adicionalmente es deseable generar un modelo que permita predecir el valor del factor más influyente en el clima a partir de los factores significativos del cual depende, utilizando para ello la técnica de regresión lineal múltiple.

2.3.3.2. Criterios de éxito para la minería de datos

Para determinar si los grupos o *clusters de* los perfiles de los colaboradores con mayor satisfacción laboral, generados por el modelo de minería de datos tengan validez, se utilizarán criterios que miden la calidad de la estructura de *clusters* generados utilizando la misma información con que se construyó el modelo. Los criterios de éxito son:

- Utilizar la suma del error al cuadrado para determinar si ha sido correcto el número correcto de grupos generado.

- Determinar si el modelo generado contiene tiene *clusters* bien definidos, comparando la suma cuadrados de error del conjunto de datos original contra un grupo de datos generados aleatoriamente que compartan la misma media y desviación estándar.

Para el modelo de regresión de la dependencia entre los factores de clima organizacional y el modelo que determine la relación entre las características demográficas del colaborador y los resultados de clima, se deberá contar con una precisión del 90%, es decir, que los modelos deben predecir correctamente el 90% de los datos de prueba.

2.4. Conocimiento de los datos

En esta fase del proyecto se examinarán de cerca los datos con los que se realizará la minería. Este paso es crítico para evitar problemas inesperados durante la siguiente fase de preparación de los datos (que comúnmente es la fase que consume mayor tiempo).

2.4.1. Recolección inicial de datos

La principal fuente de datos disponible para realizar la minería de datos es una base de datos histórica de estudios de clima organizacional realizados entre el 2010 y 2012 realizados por la empresa Estratek.

Los datos fueron entregados digitalmente en formato de copia de seguridad de la base de datos PostgreSql versión 8.3. Se procedió a cargar la copia de seguridad en una base de datos ejecutada en un servidor local, con el objetivo de describir las tablas y campos, calcular cantidad de datos, e identificar cual es la información que se va a utilizar.

2.4.1.1. Datos requeridos

Para realizar los análisis que permitan encontrar la información establecida en los objetivos de minería de datos, se debe contar con la siguiente información:

- Clasificar las evaluaciones entre las que recopilaron datos demográficos y las que no.
- Obtener las preguntas de datos demográficos que se evaluaron y las escalas nominales utilizadas en cada pregunta.

- Obtener los factores y preguntas o reactivos utilizados en las evaluaciones de clima organizacional.
- Seleccionar la información de las encuestas, los datos demográficos asociados a cada encuesta y las respuestas a cada pregunta o reactivo.
- Encontrar los índices de clima organizacional globales por evaluación o empresa, a nivel de encuesta y a nivel de factor.

2.4.1.2. Criterios de selección de datos

De la base de datos solo se utilizarán las siguientes tablas: empresa, evaluaciones, factores, preguntas, encuestas, respuestas, escalas, ítems de las escalas, rangos de evaluación.

Según información enviada por la empresa propietaria de la información, dentro de la base de datos se incluyeron evaluaciones que se utilizan de prueba y otro tipo de evaluaciones parciales que no cumplen con los lineamientos de una encuesta formal de clima organizacional por lo cual esos datos deberán ser excluidos del análisis.

Debido a que Estratek ofrece un modelo flexible, en el cual cada empresa puede establecer que factores y preguntas desea incluir en la encuesta, se debe de realizar un análisis de aquellos factores y preguntas que son comunes en la mayor parte de las evaluaciones, para tener uniformidad en los datos que se procesarán.

2.4.2. Descripción de los datos

A continuación se describen los datos proporcionados, incluyendo el formato y cantidad de los mismos, así como la identificación de tablas y campos en que se encuentra almacenada la información.

2.4.2.1. Análisis volumétrico de los datos

La base de datos contiene un total de 28 tablas, las cuales contienen un número total de 830,179 registros. Dicha información está distribuida en un total de 180 atributos o campos de las tablas.

Al realizar una primera exploración se identificaron las tablas de la base de datos que contienen la información que tiene relevancia para el estudio:

Tabla IX. **Descripción de las tablas de evaluaciones de clima organizacional**

Tabla	Descripción	# de registros
Empresa	Listado de empresas a las que se ha evaluado clima.	50
Evaluación	Listado de evaluaciones que se han realizado, una empresa puede tener más de una evaluación a través del tiempo.	71
Variable	Almacena las variables o factores de clima que se han utilizado en las mediciones.	573
Pregunta	Almacena las preguntas o reactivos que conforman las encuestas de clima.	3,857
encuesta_boleta	Representa las encuestas de clima que se han ingresado.	14,621
dato_respuesta	Contiene las respuestas asociadas a una encuesta en específico.	776,388

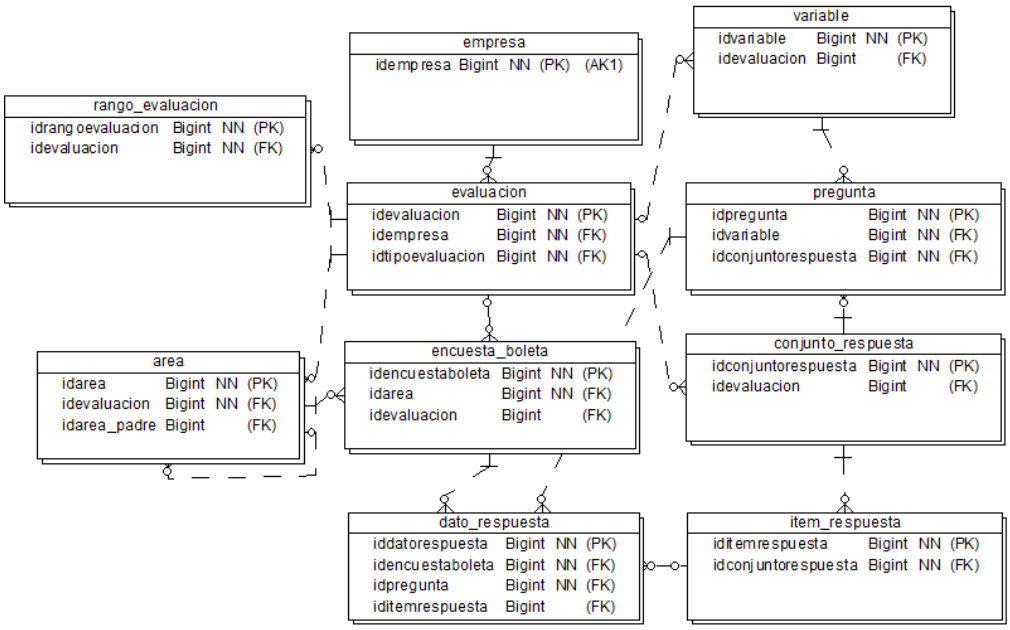
Continuación tabla IX.

conjunto_respuesta	Representa el nombre de las escalas de respuestas que se utilizan en una encuesta.	398
item_respuesta	Contiene los *ítems* de una escala de respuestas.	2,204
rango_evaluacion	Contiene los rangos por evaluación con los que se considera un clima como crítico, semicrítico o fortaleza.	213
resultado_ condensado _area	Almacena los índices de clima de las empresas y áreas que han sido evaluadas, es una tabla que condensa la información de las encuestas.	3,982

Fuente: elaboración propia.

En la figura 5 se muestra un diagrama de entidad / relación que presenta información de las tablas mencionadas y las relaciones entre ellas.

Figura 5. **Modelo relacional de la base de datos**



Fuente: elaboración propia.

2.4.2.2. Tipos de atributos y sus valores

Tipos de datos: los tipos de datos encontrados en las tablas anteriormente descritas son los siguientes:

- *"date"*: representa fecha y es utilizado únicamente para indicar la fecha en que fue realizada la evaluación.

- *"double precision"* y *"numeric"*: estos tipos de datos son utilizados indistintamente para almacenar valores numéricos flotantes. Al preparar los datos deberán unificarse en un solo tipo.

- *"integer"* y *"bigint"*: los 2 tipos representan enteros, sin embargo *bigint* soporta valores con un rango más grande. Se deberá verificar el máximo valor en este tipo de datos y de acuerdo al resultado unificarlos en un solo tipo.

- *"text"* y *"character varying"*: representan un tipo de dato que almacena caracteres alfanuméricos, deberán unificarse en un solo tipo de datos en la etapa de preparación de datos.

- *"boolean"*: almacena valores de tipo verdadero y falso, que por requerimientos de la plataforma que será utilizada para análisis, deberán convertirse a números enteros, siendo falso el valor cero y verdadero un valor uno.

La tabla X muestra los atributos relevantes de la tabla empresa de la base de datos de clima.

Tabla X. **Descripción de atributos de tabla empresa**

Atributo	Descripción	Tipo
idempresa	Llave primaria de la tabla empresa.	entero
nombre	Nombre de la empresa.	texto

Fuente: elaboración propia.

La tabla XI muestra los atributos de la tabla evaluación.

Tabla XI. **Descripción de atributos de tabla evaluación**

Atributo	Descripción	Tipo
idevaluación	Llave primaria de la tabla.	entero
idempresa	Identifica la empresa a la que pertenece la evaluación.	entero
fecha_inicio	Fecha en que se realizó el estudio.	fecha

Fuente: elaboración propia.

Al realizar el análisis de la tabla (ver tabla XII) que contiene las variables o factores que se miden en la encuesta, se identificó que por razones de diseño de la base de datos, existen nombres de variables repetidos que se identifican con diferente llave primaria debido a que pertenecen a distintas evaluaciones. Sin embargo, para los propósitos de este estudio será necesario:

- Unificar variables que tienen variaciones únicamente en la forma en que se escribieron pero representan el mismo factor de clima que se está midiendo.
- Unificar las variables que tienen el mismo nombre pero pertenecen a diferentes evaluaciones.

- En la tabla de preguntas se necesitará reasignar las relaciones hacia el nuevo listado de variables unificado.
- Quitar de este listado las filas que identifican a las variables demográficas debido a que se analizaran de otra manera.

Tabla XII. **Descripción de atributos de tabla variable**

Atributo	Descripción	Tipo
idvariable	Llave primaria de la tabla.	entero
nombre	Nombre de la variable	entero
idevaluación	Identifica la evaluación a la cual está relacionada la variable.	fecha
descripción	Contiene una descripción que explica los aspectos de clima medidos por la variable.	texto
esdato_demografico	Si el valor es verdadero indica que es una variable utilizada para identificar las características demográficas de los colaboradores.	*boolean*

Fuente: elaboración propia.

Con respecto a las preguntas de la encuesta se encuentran localizadas en la tabla “pregunta”, se debe realizar un tratamiento similar al de las variables, para unificar las preguntas equivalentes y no tener duplicada la información. En la tabla XIII se puede ver una descripción de los principales atributos de la tabla.

Tabla XIII. **Descripción de atributos de tabla pregunta**

Atributo	Descripción	Tipo
idpregunta	Llave primaria de la tabla.	Entero

Continuación tabla XIII.

idvariable	Identifica la evaluación a la cual está relacionada la variable.	entero
pregunta	Texto del reactivo con el que se pide la opinión al colaborador.	texto
idconjuntorespuesta	Identifica al tipo de escala de respuestas que utiliza esta pregunta.	entero
esdato_demografico	Si el valor es verdadero indica que es la pregunta es utilizada para identificar una característica demográfica del colaborador.	*boolean*
idtipopregunta	Si el valor es 1 es una pregunta de clima, si es 2 es demográfica, si es 3 o 4 pertenece a un tipo de pregunta complementaria que no se analizará en este estudio.	entero

Fuente: elaboración propia.

La tabla encuesta_boleta, es la que representa cada encuesta que se ha completado por parte de un colaborador, de esta tabla se toma el resultado de clima de cada colaborador en específico, dato que servirá para identificar el perfil de los colaboradores que mejor clima perciben. En la tabla XIV se identifican los principales atributos de la tabla.

Tabla XIV. **Descripción de atributos de tabla encuesta_boleta**

Atributo	Descripción	Tipo
idencuestaboleta	Llave primaria de la tabla.	entero
idevaluacion	Identifica la evaluación a la cual está relacionada la encuesta.	entero

Continuación tabla XIV.

status_tabulacion	Toma el valor 0 para una encuesta que está incompleta y 1 para una encuesta completa.	entero
resultado_evaluacion	Almacena en una escala de 0 a 100, el resultado de clima de esa encuesta en específico.	doble precisión

Fuente: elaboración propia.

Las escalas que se utilizan como opciones de respuesta en la encuesta son de tipo Likert (ver la sección del capítulo 1 llamada Instrumento de medición). Se utilizan 5 opciones de respuesta que toman los siguientes valores:

- Valor 1 si el colaborador está en total desacuerdo con la afirmación planteada.
- Valor 2 si está en desacuerdo.
- Valor 3 si está de acuerdo.
- Valor 4 si está totalmente de acuerdo.
- Valor -1 si el colaborador no puede evaluar el reactivo, si esta opción es elegida la misma es descartada para el cálculo del indicador de clima organizacional.

En algunas evaluaciones es necesario cambiar el texto de las opciones, por ejemplo, utilizar las opciones nunca, a veces, casi siempre o siempre. En la tabla XV se muestran los atributos de la tabla de datos que lista los conjuntos respuestas o escalas habilitadas para la evaluación.

Tabla XV. **Descripción de atributos de tabla conjunto_respuesta**

Atributo	Descripción	Tipo
idconjuntorespuesta	Llave primaria de la tabla.	entero
idevaluacion	Identifica la evaluación a la cual está relacionada el conjunto de respuesta.	entero
Nombre	Nombre que identifica la escala.	entero

Fuente: elaboración propia.

En la tabla XVI se describen los atributos de la tabla de datos que almacena las diferentes opciones que pertenecen a una escala de respuesta.

Tabla XVI. **Descripción de atributos de tabla item_respuesta**

Atributo	Descripción	Tipo
iditemrespuesta	Llave primaria de la tabla.	entero
idconjuntorespuesta	Identifica el conjunto de respuesta al que pertenece el item_respuesta.	entero
respuesta	Texto que se muestra para el item_respuesta en la encuesta, por ejemplo Nunca, A veces, etc.	texto
valor	Representa el número con el que se asocia el item_respuesta para posteriormente guardarlo en dato_respuesta.	entero

Fuente: elaboración propia.

La tabla dato_respuesta es de las principales tablas, ya que en ella se guardan las respuestas de los colaboradores, es acá donde se tienen que aplicar los análisis estadísticos. Sus atributos están descritos en la tabla XVII.

Tabla XVII. **Descripción de atributos de tabla dato_respuesta**

Atributo	Descripción	Tipo
iddatorespuesta	Llave primaria de la tabla.	entero
idencuestaboleta	Identifica la encuesta a la que pertenece la respuesta.	entero
idpregunta	Identifica la pregunta a la que está asociada la respuesta.	entero
valor	Valor con que se almacena la respuesta a una pregunta de la encuesta. Para las preguntas de clima toma un valor entre 1 y 4 donde 1 refleja estar en total desacuerdo con la pregunta y 4 en total acuerdo. Para las demás preguntas un valor entero positivo.	entero

Fuente: elaboración propia.

De acuerdo al modelo que utiliza la empresa Estratek, el clima organizacional de una empresa puede categorizarse en uno de los siguientes niveles: crítico, semicrítico y fortaleza, según la calificación que haya obtenido. En la tabla rango_evaluación es donde se establecen los límites de cada rango para posteriormente clasificar el resultado del clima. En la tabla XVIII se describen los atributos asociados a esa tabla de datos.

Tabla XVIII. **Descripción de atributos de tabla rango_evaluación**

Atributo	Descripción	Tipo
idrangoevaluacion	Llave primaria de la tabla.	entero
idevaluacion	Identifica la evaluación a la cual está relacionado el rango.	entero
nombre	Nombre del rango: puede ser crítico, semicrítico y fortaleza.	texto
limite_superior	Valor entre 0 y 100 que representa el límite superior del rango que se está especificando.	doble precisión

Fuente: elaboración propia.

En la tabla resultado_condensado_area (cuyos atributos se muestran en la tabla XIX) se encuentran los resultados calculados de clima organizacional a nivel de empresa y área, esta es una tabla donde se condensan los datos de las encuestas y se agrupan para poder acceder a ellos de una manera más eficiente.

Tabla XIX. **Descripción de atributos de tabla resultado_condensado_area**

Atributo	Descripción	Tipo
idvariable	Identifica la variable a la cual está relacionada el resultado condensado.	entero
idevaluacion	Identifica la evaluación a la cual está relacionada el resultado condensado.	entero
promedio	Almacena en una escala de 0 a 100, el resultado de clima de esa área, evaluación y variable en específico.	doble precisión
idarea	Identifica el área de trabajo a la cual está asociado el resultado condensado de clima.	entero

Fuente: elaboración propia.

2.4.3. Exploración de datos

Como resultado de la exploración de datos se presentan los descubrimientos que serán de utilidad para las fases posteriores de preparación y modelado de datos con el fin del alcance de los de los objetivos propuestos inicialmente:

- En total hay 61 evaluaciones descartando las que son de pruebas y las que no aplican debido a que son evaluaciones complementarias. De este número inicial hay 32 evaluaciones que incluyeron datos

demográficos y 29 restantes que no incluyeron. Estas 37 evaluaciones se utilizaran para validar y encontrar los grupos con el perfil demográfico de los colaboradores más satisfechos.

- Los datos demográficos que, de acuerdo a la información existente, se pueden analizar son los siguientes: tiempo de permanencia en la empresa, sexo, puesto de trabajo, edad, escolaridad, estado civil.

- Debido a que cada empresa puede elegir las variables que desea evaluar en el clima organizacional, existe una diferencia en las variables que se incluyen, sin embargo se logró identificar que el 85% de las evaluaciones tienen en común el 90% de las variables que miden. Esto permite tener un grupo válido para analizar la dependencia entre las variables. Las variables compartidas son las siguientes:

    - Motivación y ambiente de trabajo
    - Organización laboral
    - Liderazgo
    - Trabajo en equipo
    - Comunicación
    - Capacitación
    - Imagen e identificación

- En promedio las encuestas de clima tienen 53 preguntas. La encuesta con menos preguntas tiene 40 y la que más tiene 70.

- El promedio de preguntas por variable es de 7. Siendo el mínimo 3 preguntas por variable y el máximo 13.

- La tabla de encuesta_boleta posee un atributo importante: el resultado del clima del colaborador que completo la encuesta, en la tabla XX se puede ver los principales estadísticos descriptivos para la variable. De esta estadística se resalta el hecho de que hay un valor máximo de 104 cuando el índice de clima organizacional solo puede tener un rango de 25 a 100 puntos, lo que indica que hay un error en el cálculo de algunas filas que habrá que tomar en cuenta en la limpieza de datos.

Tabla XX. **Estadística descriptiva de resultados de clima organizacional**

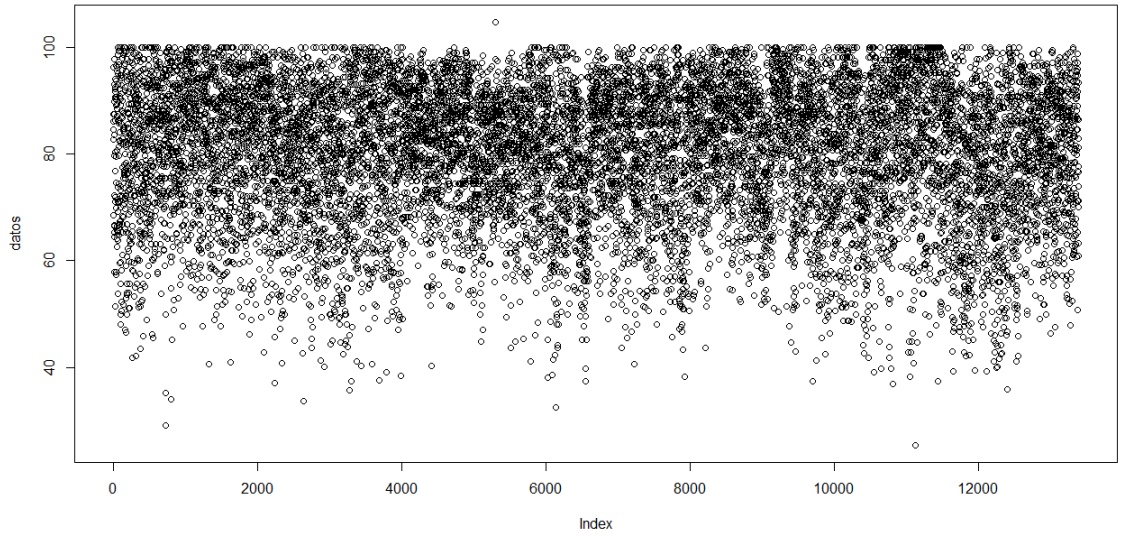
Estadístico	Valor
Mínimo	25.47
Máximo	104.7
Media	80.23
Desviación estándar	12.58
1er Cuartil	72.00
Mediana	82.05
3er Cuartil	90.16

Fuente: elaboración propia.

- En la figura 6 se muestra una gráfica que representa la distribución de resultados de clima organizacional de las 13,397 encuestas que se seleccionaron de la base de datos. La gráfica presenta en el eje “X” el número de dato que se está graficando y en el eje “Y” el índice de satisfacción de clima organizacional de cada encuesta. Como se puede observar los resultados de clima oscilan entre 25 y 100 puntos. El área con mayor densidad de datos está entre 80 y 90 puntos lo que se alinea con la media de 80 puntos mostrada en la tabla XX. También se puede observar que hay una encuesta por arriba de 100, lo que indica que es un dato erróneo debido a que el máximo punteo posible es de 100, por lo

cual habrá que verificar la validez de la encuesta que representa ese dato.

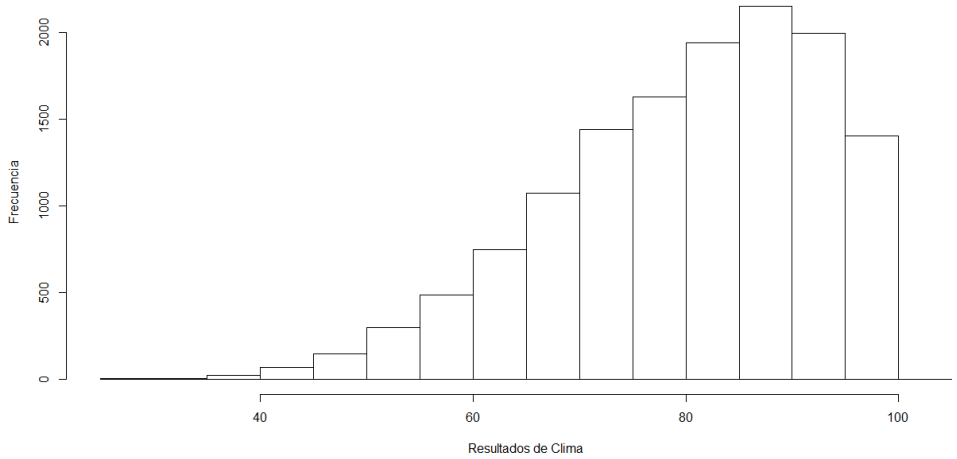
Figura 6. **Distribución de resultados de clima por colaborador**



Fuente: elaboración propia, generada con el software R.

- En la figura 7 se muestra un histograma con los resultados de clima organizacional. Los datos fueron agrupados en rangos de 5 puntos para identificar las calificaciones de clima organizacional más frecuentes, donde se observa que entre 80 y 95 puntos son los más frecuentes.

Figura 7. **Histograma de resultados de clima por colaborador**



Fuente: elaboración propia, generado con el software R.

2.4.4. Verificar calidad de los datos

Para este proyecto en específico los datos que se utilizarán en la minería provienen de una base de datos relacional implementada en PostgreSQL, lo cual proporciona la ventaja de que los datos siguen un modelo relacional que garantiza que los datos estén estructurados y normalizados. Adicionalmente el software *R* tiene paquetes que permiten acceder directamente a una base de datos de PostgreSQL lo que hace eficiente la obtención de datos y evita el uso de formatos de transferencia (como csv, xml, etc) que añaden complejidad e inducen a una mayor probabilidad de error en la migración de la información.

Aunque las consideraciones descritas en el párrafo anterior resultan en una ventaja para el estudio, no garantizan un conjunto de datos libre de errores. A continuación se detallan algunos aspectos encontrados, los cuales se deberán corregir, completar o considerar en la próxima fase de preparación de los datos.

2.4.4.1. Datos faltantes

A las encuestas de clima organizacional que son completadas electrónicamente en el sistema de Estratek se les pone una restricción para que todas las preguntas sean contestadas, sin embargo en las encuestas que son aplicadas físicamente existe la posibilidad de que el colaborador no haya contestado todas las preguntas, es por eso que en la tabla dato_respuesta se encuentran valores nulos o bien filas faltantes debido a que no se registró la respuesta. El número de encuestas con respuestas faltantes es de 222, lo que representa menos del 2% del total de encuestas.

Otra información que se debe completar son los atributos del listado de empresas evaluadas que se indicaron en la sección 2.4.1.1.

2.4.4.2. Datos incorrectos

Cuando se realizó la estadística descriptiva de la tabla encuesta_boleta para analizar el atributo de resultado clima organizacional se identificó una encuesta con un resultado de 104.68, sin embargo, solo es posible tener un resultado entre 25 y 100 puntos, por lo que hay que revisar, corregir o anular las encuestas que presenten este tipo de dato incorrecto.

En la tabla de datos_respuesta que recopila las respuestas codificadas a las preguntas de las encuestas, se encontraron 5 datos que tienen valores fuera del rango permitido para la escala, es decir, valores diferentes de -1, 1, 2, 3 y 4. Estos se deberán eliminar del conjunto de datos a analizar.

Hay descripciones de preguntas y variables que poseen errores de tipo ortográfico que serán corregidos en la fase de preparación de la información.

2.4.4.3. Inconsistencia en las codificaciones

Como los datos demográficos se especifican de manera independiente para cada evaluación, el valor codificado que se utilizan para identificar cada elemento puede variar. Por ejemplo para una evaluación el rango de edad de 18 a 22 años se codifica con el valor 1, mientras que para otra evaluación se codifica con el valor 2. Una de las tareas de la preparación de datos será alinear las codificaciones de los datos demográficos.

Con las variables y preguntas sucede lo mismo que con los datos demográficos, aunque se refiera a la misma variable o pregunta, se utiliza un identificador diferente, por lo cual se tendrá que crear una nueva codificación y alinear los valores de todas las evaluaciones con la codificación definida.

2.5. Preparación de los datos

La preparación de datos es una de las fases más importante de la minería de datos. Se realizan tareas como la selección de datos, mezclar datos de diversas fuentes, crear nuevos atributos con los datos existentes y reemplazar valores nulos o en blanco del conjunto original de datos.

2.5.1. Selección de datos

Para el análisis del perfil demográfico de los colaboradores se seleccionaron únicamente aquellas evaluaciones que incluyeron la captura de datos demográficos en las encuestas. A continuación se describe la selección del conjunto de datos:

- 32 evaluaciones de clima organizacional realizadas entre los años 2010 y finales del 2012.
- Esto representa un total de 10,400 colaboradores encuestados de 25 diferentes empresas.
- Los atributos que se recopilarán son las características demográficas de cada colaborador y el resultado de clima organizacional para cada colaborador.

La selección de datos para el segundo objetivo del presente análisis en el cual se plantea descubrir la dependencia o entre los factores de clima organizacional, se basa en el siguiente conjunto de datos:

- Se utilizarán 50 evaluaciones, las cuales se caracterizan por tener en común la medición de 7 factores de clima organizacional. Las 11 descartadas carecían de 3 a 4 de los 7 factores mostrados en la tabla XXI. Los 7 factores que se incluyen en el análisis son los descritos en la tabla XXI.

Tabla XXI. **Factores que se analizarán en clima organizacional**

Factores de clima organizacional
Capacitación y desarrollo
Comunicación
Identificación e imagen
Liderazgo
Motivación y ambiente de trabajo
Organización y cambio
Trabajo en equipo

Fuente: elaboración propia.

- El total de encuestas aplicadas a colaboradores que se tomarán en cuenta para este conjunto de datos es de 9,153.

2.5.2. Limpieza de datos

En la fase de la exploración de datos se identificaron varios aspectos y datos que deben ser corregidos o eliminados del conjunto de datos final, porque contenían problemas de calidad. A continuación se describen los procesos que se realizaron para garantizar la calidad de la información analizada.

2.5.2.1. Datos faltantes

En la tabla *dato_respuesta* se encuentran filas con valores nulos debido a que no se registró la respuesta en el proceso de llenado de encuestas. Como se indicó estos valores representan menos del 2% del total de encuestas, por lo que sin tener un impacto negativo en la proporción de la información se eliminaron las filas con valores nulos. En este caso no hay forma de completar está información porque es la que el colaborador responde, de otra manera se estaría sesgando los datos.

En la tabla XXII se presentan el conteo de los datos después de realizar la limpieza de datos nulos.

Tabla XXII. **Conteo de datos válidos en la tabla dato_respuesta**

Descripción	Número de Filas
Total de filas en la tabla dato_respuesta.	33

Continuación tabla XXII

Filas con valores nulos (que serán removidas)	1
Total de filas con valores completos	32

Fuente: elaboración propia.

2.5.2.2. Errores en los datos

En la tabla encuesta_boleta se reportó un dato de promedio de clima para una encuesta cuyo es de 104.68 que sobrepasa el máximo valor de 100 que puede generarse en estos casos. Se consultó con la empresa Estratek y reporta que se trata de un error aislado y que las demás encuestas de esa evaluación pueden ser tomadas en cuenta debido a que se verificaron los valores y son correctos. Debido a esto se procedió a eliminar esa encuesta del conjunto de datos final.

En la misma evaluación descrita en el párrafo anterior, se encontraron 5 datos que tienen valores fuera del rango permitido para la escala, es decir, valores diferentes de -1, 1, 2, 3 y 4. Estos se deben a una tabulación incorrecta de los datos, los mismos ya no pueden ser recuperados por lo que se procederá a eliminar las 5 respuestas que pertenecen a 5 encuestas de diferentes colaboradores.

2.5.2.3. Inconsistencia en las codificaciones

Los datos demográficos presentes en las base de datos tienen inconsistencia en la codificación, debido a que cada evaluación maneja sus propios códigos los cuales son almacenados en la información de la encuesta.

Por lo cual se propuso una codificación única para cada tipo de dato demográfico, en las siguientes tablas se presentan las codificaciones utilizadas en el conjunto de datos final.

Tabla XXIII. **Codificación para tiempo de permanencia del colaborador**

Tiempo de Permanencia	Código
0 a 2 años	1
2 a 5 años	2
5 a 10 años	3
10 a 15 años	4
más de 15 años	5

Fuente: elaboración propia.

Tabla XXIV. **Codificación para género del colaborador**

Género	Código
Femenino	1
Masculino	2

Fuente: elaboración propia.

Tabla XXV. **Codificación para puesto de trabajo del colaborador**

Puesto de Trabajo	Código
Gerencial	1
Jefatura	2
Administrativo	3
Operativo	4

Fuente: elaboración propia.

Tabla XXVI. **Codificación para rango de edad del colaborador**

Rango de Edad	Código
18 a 25 años	1
26 a 30 años	2
31 a 40 años	3
41 a 50 años	4
más de 50 años	5

Fuente: elaboración propia.

Tabla XXVII. **Codificación para escolaridad del colaborador**

Escolaridad	Código
Sin educación escolar	1
Primaria	2
Básicos	3
Diversificado	4
Licenciatura	5
Posgrado	6

Fuente: elaboración propia.

Tabla XXVIII. **Codificación para estado civil del colaborador**

Estado Civil	Código
Soltero	1
Casado / Unido	2

Fuente: elaboración propia.

2.5.3. Construcción de datos

La información necesaria para ejecutar el algoritmo de *clustering* debió ser construida a partir de la información recabada en varias tablas y un proceso de reasignación de códigos:

- Debido a que los datos demográficos de una encuesta se encuentran almacenados como preguntas dentro de la encuesta, debieron migrarse a una tabla nueva que poseía la información del identificador de la encuesta, el resultado de clima y el dato demográfico al que pertenece.

- Posteriormente se extrajo un resumen de todos los tipos de datos demográficos en otra tabla, a partir de esa tabla resumen se reasignaron los códigos de acuerdo a las tablas de codificación mostradas en el numeral anterior.

- Finalmente se creó una tabla de datos con la información de acuerdo al formato que se especifica en la tabla XXXI. En este proceso se encontró información de datos demográficos que no forman parte del listado que se seleccionó, por lo que debieron ser eliminadas las encuestas, ya que no proporcionan ningún valor para el análisis. El número de encuestas del conjunto final es de 7,487. En la tabla XXIX se muestra la lista de datos demográficos y el número de encuestas que poseen información del mismo.

Tabla XXIX. **Número de registros asociados a los datos demográficos**

Dato Demográfico	Número de Registros
Tiempo de permanencia	7046
Puesto de trabajo	320
Estado civil	7197
Escolaridad	7299
Edad	7295
Género	7344

Fuente: elaboración propia.

El proceso de construcción de datos, para el análisis de regresión entre factores de clima organizacional, fue más corto, esto debido a que en la tabla *resultado_condensado_area* proporcionada por Estratek, ya se encontraba consolidados los resultados de clima por factor y por evaluación. Las tareas realizadas en la construcción de estos datos fueron las siguientes:

- Se generó un listado de todas las evaluaciones realizadas, que factores de clima se incluyeron y cuál fue el punteo obtenido.

- Se seleccionaron las evaluaciones que evaluaron los 7 factores definidos en la tabla XXI.

- Los datos de estas evaluaciones seleccionadas se prepararon de acuerdo al formato propuesto en la tabla XXXI.

2.5.4. Dar formato a los datos

Antes de pasar a la etapa de modelado, es importante que los datos estén preparados en el formato requerido por la herramienta de modelado. Para la investigación fue necesario preparar los datos para los siguientes formatos de acuerdo a la técnica de minería que se aplicará.

2.5.4.1. *Clustering* o segmentación de datos

Para el análisis de los grupos demográficos con mejor clima organizacional se utilizó el algoritmo *k-means.* El programa *R* requiere que los datos se ingresen con el formato presentado en la tabla XXX.

Tabla XXX. **Formato de datos requerido para el *clustering***

Objeto	Atributo 1	Atributo 2	Atributo 3
A	1	1	2
B	2	1	3
C	4	3	2
D	5	4	1

Fuente: elaboración propia.

Este formato presenta los objetos u observaciones en las filas y cada uno de los atributos o variables se deben incluir como columnas. Los nombres de los objetos no pueden repetirse, los valores que identifican los atributos deben de ser de tipo numérico.

Adaptando este formato a los datos de clima organizacional, la tabla de datos quedará con los encabezados que se muestran en la tabla XXXI.

Tabla XXXI. **Formato para *clustering* de clima organizacional**

TIPO	DATO DEMOGRÁFICO	CAPACITACIÓN	COMUNICACIÓN	IDENTIFICACIÓN E IMAGEN
Edad	18 a 25 años	77.65	76.17	85.07
Escolaridad	Diversificado	73.63	75.88	84.01
Estado Civil	Soltero	77.53	75.92	84.33

Fuente: elaboración propia.

2.5.4.2. Regresión lineal múltiple

La técnica que permitió establecer si existe dependencia entre las variables medidas en el clima organizacional es la regresión lineal múltiple. Este modelo requiere que los datos estén dispuestos en una matriz, donde cada columna representa un factor o variable de clima organizacional y cada fila representa una evaluación y su índice de satisfacción con respecto al factor indicado por el encabezado de la columna. Un ejemplo del formato de entrada para el algoritmo se muestra en la tabla XXXII.

Tabla XXXII. **Formato de entrada del modelo de regresión lineal múltiple de los factores de clima organizacional**

Capacitación	Comunicación	Identificación	Liderazgo	Motivación	Organización
3.3586	3.2391	3.5712	3.3245	3.3076	3.3459
3.3240	3.3071	3.7222	3.3979	3.5114	3.5323
3.0963	3.0057	3.4831	3.2817	3.2096	3.2190
3.3929	3.1584	3.6746	3.3005	3.4429	3.3761
3.2456	2.9323	3.5263	2.9585	3.3128	3.3487
2.9010	2.9071	3.4649	3.0533	3.2487	3.3659
3.3763	3.4093	3.5725	3.6004	3.2866	3.5886
2.9463	2.9415	3.4038	3.1440	3.0773	3.2599

Fuente: elaboración propia.

2.6. Modelado

Esta etapa se construyen los modelos de datos que permiten descubrir el conocimiento implícito dentro de los datos.

2.6.1. Selección de técnicas de modelado

Debido a que se han establecido 2 objetivos principales de minería de datos, se realizó un análisis para identificar cual es la técnica más apropiada para cada objetivo.

2.6.1.1. Modelo de *clustering* o segmentación

La segmentación tiene como objetivo la separación de los datos en subgrupos o clases de alto interés y significado. Este proceso puede realizarse de manera manual o semi-automática. Uno de los objetivos del presente trabajo es encontrar los grupos demográficos que tengan un mejor clima organizacional dentro de sus empresas, por lo que se seleccionó una técnica de segmentación para construir el modelo de minería de datos.

Específicamente se utilizó un algoritmo ampliamente conocido para realizar *clustering* llamado *k-means,* el cual consiste en un análisis divisivo y no jerárquico para segmentar grupos. Las características y requisitos para la utilización del algoritmo *k-means* son:

- Como es un algoritmo no supervisado, no se necesitan subconjuntos de datos de entrenamiento y de prueba.

- Los datos deben presentarse en una matriz que contenga una columna con las observaciones y las demás columnas, las características de las observaciones, que en este caso son los datos demográficos.

- Ninguna de las filas puede tener atributos vacíos, todas las observaciones deben tener valores asociados a cada atributo.

2.6.1.2. Modelo de regresión lineal múltiple

Con el objetivo de determinar si existe dependencia entre las distintas variables del clima organizacional, se decidió utilizar el método de regresión lineal múltiple aplicado a las mediciones de clima.

La regresión lineal múltiple permite analizar conjuntos de datos que contengan 2 o más variables y determinar si existe dependencia de una variable sobre las demás, permitiendo encontrar relaciones de causa y efecto entre las variables de clima organizacional. Esto se realiza por medio de la construcción de un modelo lineal matemático, que permita predecir los valores de la variable dependiente.

Para establecer cuál de los factores de clima organizacional será elegido como variable dependiente del modelo de regresión, se realizó un análisis de correlación entre las distintas variables y el índice de clima organizacional y se seleccionó la variable más influyente, para saber si esta a su vez depende de otras.

2.6.2. Construir modelo de c*lustering*

Para construir el modelo de *clustering* se utilizó el algoritmo *k-means* implementado en el software R en la librería llamada *cluster*. Las observaciones que serán agrupadas son los segmentos demográficos y las características que los describen son las calificaciones de cada una de las variables de clima organizacional.

2.6.2.1. Configuración de parámetros

Para la ejecución del algoritmo *k-means* es necesario especificar los siguientes parámetros:

- Datos: una matriz conformada por valores numéricos con los datos que se analizarán. Para este caso específico, tiene el siguiente formato:

Tabla XXXIII. **Parámetro de entrada de datos para algoritmo *k-means***

TIPO	DATO DEMOGRÁFICO	CAPACITACIÓN	COMUNICACIÓN	IDENTIFICACIÓN E IMAGEN
Edad	18 a 25 años	77.65	76.17	85.07
Edad	26 a 30 años	77.41	76.49	83.95
Escolaridad	Diversificado	73.63	75.88	84.01
Estado Civil	Casado / Unido	77.03	75.41	84.49
Estado Civil	Soltero	77.53	75.92	84.33
Género	Femenino	77.70	77.94	85.55
Tiempo de Permanencia	2 a 5 años	78.83	76.36	84.94

Fuente: elaboración propia.

- Centros o número de *clusters*: este parámetro puede tomar 2 valores, si es un entero, representa el número de grupos que se quieren obtener como resultado. O bien, puede ser una matriz con los centros con que se desea iniciar el algoritmo. Para el modelo que se está proponiendo el número de *clusters* elegido es 3. Más adelante en el apartado Evaluación del Modelo se especifican los criterios utilizados para elegir el valor de este parámetro.

- Número máximo de iteraciones: del algoritmo que se puede ejecutar. Se pudo observar que más allá de 20 iteraciones los grupos conformados como resultado del algoritmo no variaban. Por lo cual se estableció como número máximo de iteraciones permitidas 20.

- Conjuntos aleatorios: si en el segundo parámetro no se especificaron los centros, sino el número de *clusters*, se debe indicar el número de datos aleatorios que se tomaran para elegir los centros iniciales**.** Para este caso se elegirán 15 conjuntos aleatorios.

Tabla XXXIV. **Tabla de parámetros de configuración del algoritmo *k-means***

Parámetro	Valor
Datos	Datos del modelo
Número de *clusters*	3
Número máximo de iteraciones	20
Número de conjuntos aleatorios utilizados para los centros	15

Fuente: elaboración propia.

2.6.2.2. Descripción del modelo

A partir de la ejecución del algoritmo con los parámetros indicados en el numeral anterior, se obtuvo un modelo de *clusters* con las siguientes características:

- *Clusters*: los distintos datos demográficos fueron agrupados en 3 *clusters* que se muestran en la tabla XXXV.

Tabla XXXV. ***Clusters* conformados**

TIPO	DATO DEMOGRÁFICO	*CLUSTER*
Edad	18 a 25 años	1
Edad	26 a 30 años	1
Escolaridad	Diversificado	1
Estado civil	Casado / Unido	1
Estado civil	Soltero	1
Género	Femenino	1
Tiempo de permanencia	2 a 5 años	1
Edad	más de 50 años	2
Escolaridad	Posgrado	2
Escolaridad	Primaria	2
Puesto	Administrativo	2
Puesto	Gerencial	2
Puesto	Jefatura	2
Puesto	Operativo	2
Tiempo de permanencia	10 a 15 años	2
Tiempo de permanencia	más de 15 años	2
Edad	31 a 40 años	3
Edad	41 a 50 años	3
Escolaridad	Básicos	3
Escolaridad	Licenciatura	3
Género	Masculino	3
Tiempo de permanencia	0 a 2 años	3
Tiempo de permanencia	5 a 10 años	3

Fuente: elaboración propia.

- Centros: los centros de cada uno de los *clusters* son los siguientes:

Tabla XXXVI. **Centros de los *clusters* generados**

	CLUSTER		
VARIABLE	**1**	**2**	**3**
CAPACITACIÓN	77.11	81.59	79.76
COMUNICACIÓN	76.31	80.39	78.32
IDENTIFICACIÓN E IMAGEN	84.62	89.28	87.21
LIDERAZGO	77.31	81.45	79.09
MOTIVACIÓN Y AMBIENTE	79.80	84.06	82.17
ORGANIZACIÓN Y CAMBIO	79.98	84.86	82.01
TRABAJO EN EQUIPO	74.98	80.65	78.36

Fuente: elaboración propia.

- Tamaño: el tamaño de cada *cluster* quedo conformado de la siguiente manera:

Tabla XXXVII. **Tamaños de *cluster***

Cluster	Número de elementos	Porcentaje
1	7	30.4%
2	9	39.1%
3	7	30.4%

Fuente: elaboración propia.

- Las sumas de cuadrados dentro de cada *cluster*: este dato indica suma de las distancias al cuadrado entre los puntos del *cluster* y su centro (cuyos componentes son las medias de las distintas variables).

Tabla XXXVIII. **Suma de cuadrados dentro de cada *cluster***

Cluster	Suma de Cuadrados dentro de cada *cluster*
1	72.5175
2	77.4824
3	158.4349
Total	**308.4348**

Fuente: elaboración propia.

- El total de la suma de cuadrados dentro de cada *cluster*: 902.973.

- Suma de cuadrados de la distancia entre *clusters*: es de 594.54.

2.6.2.3. Interpretación de resultados

Como resultado de la aplicación del algoritmo *k-means* el conjunto de datos inicial fue segmentado en 3 grupos o *clusters*. Para la interpretación se tomó en cuenta el índice de clima organizacional para cada variable y el segmento demográfico del grupo.

En las tablas XXXIX, XL y XLI se muestran las características demográficas clasificadas, los promedios del índice de clima organizacional para cada variable, las cuales tienen un sombreado de varios colores y matices donde rojo indica un bajo índice de clima, blanco un índice promedio y azul un índice alto.

Para la interpretación de los grupos debe tomarse en cuenta que cada segmento demográfico está incluido de manera independiente, por ejemplo en el grupo 1, están los segmentos demográficos de rango de edad de 18 a 25

años y escolaridad: diversificado, lo cual no implica que en ese grupo hayan sido tomados colaboradores de 18 a 25 años que hayan alcanzado el grado de diversificado, más bien que hay grupos de colaboradores de 18 a 25 años y que también está el grupo de colaboradores con escolaridad a nivel diversificado.

- *Cluster* 1: muestra los grupos demográficos con el índice más bajo de clima organizacional.

  Características demográficas
  - Edad: los colaboradores están en edades de 18 a 30 años
  - Tiempo de permanencia en la empresa: 2 a 5 años
  - Estado civil: casado o unido y soltero
  - Género: femenino

  Variables de clima con menor satisfacción: capacitación, comunicación, liderazgo y trabajo en equipo.

De las características de este grupo se puede concluir que lo integran colaboradores que aún están en proceso de formación profesional y académica, lo cual denota una mayor insatisfacción con respecto a la capacitación que reciben o la falta de ella.

Este grupo no percibe un buen liderazgo, el cual puede estar influenciado por una mala comunicación.

La característica demográfica de estado civil no es representativa del grupo, debido a que en el mismo aparecen todos los segmentos posibles para esa característica: casado o unido y soltero.

Tabla XXXIX. ***Cluster* 1 con clima organizacional bajo**

TIPO	DATO DEMOGRAFICO	CLUSTER	CAPACITACION	COMUNICACIÓN	IDENTIFICACION E IMAGEN	LIDERAZGO
Edad	18 a 25 años	1	77.65	76.17	85.07	79.84
Edad	26 a 30 años	1	77.41	76.49	83.95	78.64
Escolaridad	Diversificado	1	73.63	75.88	84.01	74.76
Estado Civil	Casado / Unido	1	77.03	75.41	84.49	73.62
Estado Civil	Soltero	1	77.53	75.92	84.33	77.70
Género	Femenino	1	77.70	77.94	85.55	79.66
Tiempo de Permanencia	2 a 5 años	1	78.83	76.36	84.94	76.96

TIPO	DATO DEMOGRAFICO	CLUSTER	MOTIVACION Y AMBIENTE	ORGANIZACIÓN Y CAMBIO	TRABAJO EN EQUIPO	PROMEDIO
Edad	18 a 25 años	1	79.95	79.15	77.29	79.30
Edad	26 a 30 años	1	78.59	79.85	75.46	78.63
Escolaridad	Diversificado	1	79.75	79.31	73.28	77.23
Estado Civil	Casado / Unido	1	78.97	79.36	74.30	77.60
Estado Civil	Soltero	1	79.62	80.23	74.62	78.57
Género	Femenino	1	81.34	81.48	73.63	79.62
Tiempo de Permanencia	2 a 5 años	1	80.37	80.44	76.25	79.16

Fuente: elaboración propia.

- *Cluster* 2: en este *cluster* quedaron representados los colaboradores con el mejor índice de clima organizacional en comparación con los otros 2 *clusters*.

Características demográficas:

- Edad: más de 50 años
- Tiempo de permanencia en la empresa: de 10 años en adelante
- Puesto: administrativo, gerencial, jefatura, operativo
- Escolaridad: posgrado y primaria

Fortalezas en clima organizacional: identificación e imagen, motivación y ambiente de trabajo, organización laboral y cambio.

Las características de este grupo reflejan que son los colaboradores con más madurez y con una alta estabilidad laboral, a su vez son los que tienen una mayor identificación con la empresa. Otra característica del grupo, es que poseen un alto nivel de motivación y que ven positivamente el tema de organizacional laboral dentro de la empresa.

Del hecho de que en este grupo estén los 5 segmentos de los puestos de trabajo, se puede concluir que esta característica no diferencia al grupo

.

Tabla XL. ***Cluster* 2 con clima organizacional alto**

TIPO	DATO DEMOGRAFICO	CLUSTER	CAPACITACION	COMUNICACIÓN	IDENTIFICACION E IMAGEN	LIDERAZGO
Edad	mas de 50 años	2	80.21	80.81	90.50	80.46
Escolaridad	Postgrado	2	81.60	80.07	87.61	83.89
Escolaridad	Primaria	2	81.13	80.42	88.22	80.99
Puesto	Administrativo	2	81.59	80.65	89.73	85.51
Puesto	Gerencial	2	85.66	79.60	90.00	77.86
Puesto	Jefatura	2	81.11	80.62	90.99	83.18
Puesto	Operativo	2	79.95	79.95	88.46	80.09
Tiempo de Permanencia	10 a 15 años	2	82.23	80.64	89.65	81.01
Tiempo de Permanencia	mas de 15 años	2	80.85	80.78	88.39	80.08

TIPO	DATO DEMOGRAFICO	CLUSTER	MOTIVACION Y AMBIENTE	ORGANIZACIÓN Y CAMBIO	TRABAJO EN EQUIPO	PROMEDIO
Edad	mas de 50 años	2	86.32	85.15	82.59	83.72
Escolaridad	Postgrado	2	84.33	83.23	78.65	82.77
Escolaridad	Primaria	2	85.51	83.73	83.05	83.29
Puesto	Administrativo	2	84.01	86.11	80.77	84.05
Puesto	Gerencial	2	81.25	87.28	79.46	83.02
Puesto	Jefatura	2	83.65	84.96	80.26	83.54
Puesto	Operativo	2	83.65	84.12	81.20	82.49
Tiempo de Permanencia	10 a 15 años	2	86.87	84.18	76.91	83.07
Tiempo de Permanencia	mas de 15 años	2	80.95	85.00	82.96	82.71

Fuente: elaboración propia.

- *Cluster* 3: muestra los grupos demográficos con un índice de clima promedio.

  Características demográficas:

  - Edad: 31 a 50 años
  - Escolaridad: básicos y licenciatura
  - Género: masculino
  - Tiempo de permanencia en la empresa: de 0 a 2 años y de 5 a 10 años:

  Este grupo que percibe el clima laboral de una manera aceptable, está conformado por colaboradores de 30 a 50 años, que bien estén iniciando un nuevo trabajo (0 a 2 años de permanencia) o estén en plena carrera profesional dentro de la empresa (5 a 10 años).

Tabla XLI. ***Cluster* 3 con clima organizacional promedio**

TIPO	DATO DEMOGRAFICO	CLUSTER	CAPACITACION	COMUNICACIÓN	IDENTIFICACION E IMAGEN	LIDERAZGO
Edad	31 a 40 años	3	80.12	79.13	87.76	79.84
Edad	41 a 50 años	3	81.87	80.24	88.04	80.25
Escolaridad	Basicos	3	78.85	76.59	88.22	75.31
Escolaridad	Licenciatura	3	77.29	77.65	86.89	80.18
Género	Masculino	3	79.64	78.16	85.66	78.39
Tiempo de Permanencia	0 a 2 años	3	80.98	78.47	86.10	80.58
Tiempo de Permanencia	5 a 10 años	3	79.55	78.03	87.79	79.09

TIPO	DATO DEMOGRAFICO	CLUSTER	MOTIVACION Y AMBIENTE	ORGANIZACIÓN Y CAMBIO	TRABAJO EN EQUIPO	PROMEDIO
Edad	31 a 40 años	3	82.66	81.42	78.83	81.39
Edad	41 a 50 años	3	81.97	83.05	76.90	81.76
Escolaridad	Basicos	3	84.43	81.32	81.18	80.84
Escolaridad	Licenciatura	3	81.07	82.38	76.73	80.31
Género	Masculino	3	81.03	81.85	78.31	80.43
Tiempo de Permanencia	0 a 2 años	3	81.31	81.48	77.69	80.94
Tiempo de Permanencia	5 a 10 años	3	82.74	82.55	78.87	81.23

Fuente: elaboración propia.

2.6.2.4. Evaluar modelo

Para la validación y construcción del modelo, uno de los parámetros más importantes a determinar fue el número de *clusters* que se deseaban obtener a partir del conjunto de datos inicial.

En términos generales, el criterio utilizado para determinar cuál es el número óptimo de *clusters,* consiste en encontrar un balance entre la máxima compresión de los datos utilizando un único *cluster* (mayor porcentaje de error, pero mejor agrupamiento) y la máxima precisión utilizando un *cluster* por cada elemento de datos (cero porcentaje de error con nulo agrupamiento).

En este trabajo se realizó un análisis de varios modelos de *clustering,* los cuales utilizaban el mismo conjunto de datos y se variaba el número de *clusters* para cada modelo. De cada modelo generado se registró la suma de cuadrado del error, la cual se define como la suma del cuadrado de la distancia que existe entre cada elemento del *cluster* y el centro del *cluster*. La suma de cuadrados representa una medida global del error del modelo. A medida que incrementa el número de *clusters*, la suma de cuadrados se hace más pequeña, ya que el error disminuye. [12]

Al graficar la suma de cuadrados versus una serie secuencial de número de *clusters*, se observa de forma gráfica la manera de elegir el número apropiado número de *clusters*. Para ello es necesario identificar en que punto de la gráfica la suma de cuadrados se reduce drásticamente, para ello se debe buscar una sección en “forma de codo” en la gráfica, lo que indica que a partir

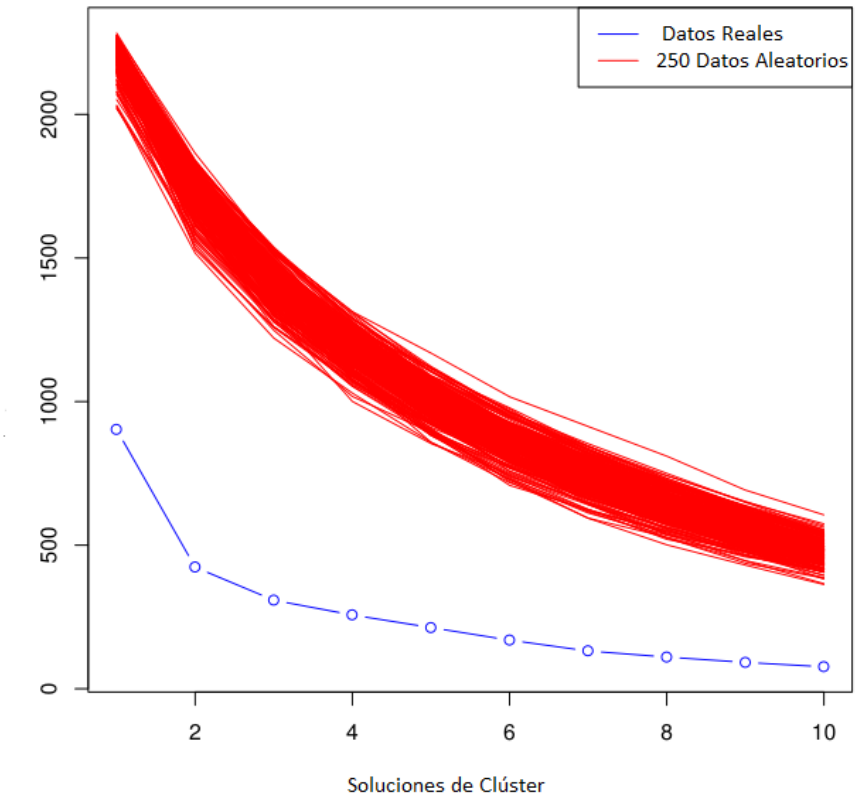
[12] PEEPLES, Matthew. *R Script for K-Means Cluster Analysis.* [en línea] <http://www.mattpeeples.net/kmeans.html> [Consulta: 5 de julio de 2013].

de ese número de *clusters* el error total del modelo no disminuirá significativamente.

En la figura 8 se muestra la gráfica del análisis de suma de cuadrados, para el modelo de *clusters* de datos demográficos de clima organizacional. Para generar está gráfica se generaron modelos alternos, con soluciones que utilizan desde 1 *cluster* hasta soluciones con 10 *clusters*.

Como validación adicional y con la ayuda de un *script* automatizado se generaron 250 versiones del conjunto de datos original a los cuales se les calculó la suma de cuadrados. Estas 250 versiones tienen datos generados aleatoriamente para cada variable, que comparten la misma media y desviación estándar que los datos originales. Si el conjunto de datos original tiene *clusters* bien definidos, la suma de los cuadrados debe reducirse más rápidamente que en los conjuntos de datos aleatorios a medida que se incrementa el número de *clusters*.

Figura 8. **Gráfica de soluciones de *clusters* versus suma de cuadrados**



Fuente: elaboración propia, generada con el programa R.

Como primera conclusión de la gráfica mostrada en la figura 8, se observa que “el codo” se visualiza al utilizar una solución de 2 *clusters*. Incluso hay una disminución significativa al utilizar 3 *clusters*. Para el desarrollo del modelo de este trabajo se tomó la decisión de utilizar un modelo de 3 *clusters*, al comparar el modelo de 2 contra el de 3 *clusters*, se identificó que el de 3 ofrece una mejor categorización de los elementos demográficos, permitiendo tener un grupo que califica como alto el clima organizacional, uno promedio y el otro bajo.

Otra conclusión que se puede obtener de la figura 8, es que al contrastar los datos del modelo original versus los datos aleatorios, el conjunto de datos

original tiene *clusters* bien definidos. Lo cual se comprueba al observar que la suma de cuadrados de error decrece más rápido para el conjunto original, que en el caso de los datos aleatorios.

2.6.3. Construir modelo de regresión lineal múltiple

El desarrollo del modelo de regresión se realizó en el software *R*. En este caso se ejecutó la función llamada *lm,* la cual es utilizada para ajustar modelos lineales, llevar a cabo la regresión y obtener datos de análisis de la varianza y covariancia del modelo.

Cuando se quiere construir un modelo de regresión se debe formular la hipótesis de cuál es la variable dependiente. En el caso de esta investigación es de alto interés conocer si hay algún factor que influye más en el índice de clima organizacional. Es por ello que primero se determinó cual es la variable más influyente en los resultados de clima, lo que permitirá orientar eficazmente los planes de acción para la mejora de clima.

2.6.3.1. Determinando la variable dependiente a utilizar

Para ello se analizó la correlación entre las variables de clima. "La correlación desde el punto de vista estadístico indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas. Se puede decir que dos variables están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra."[13]

[13] *Correlación.* [en línea] http://es.wikipedia.org/wiki/Correlaci%C3%B3n. [Consulta: 10 de julio de 2013].

Se utilizó el coeficiente de correlación de Pearson que proporciona una medida de relación lineal entre las variables y tiene la característica de ser independiente de la escala de medidas que tengan las variables. Para interpretar el resultado de correlación se utilizó la guía descrita en la tabla XLII.

Tabla XLII. **Guía para interpretar índice de correlación de Pearson**

Índice de Correlación: r	Interpretación
r = 1	Existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
0 < r < 1	Existe una correlación positiva. Si r es muy cercano a 1, indica un alto nivel de correlación.
r = 0	No existe relación lineal.
-1 < r < 0	Existe una correlación negativa. Si r es muy cercano a -1, indica un alto nivel de correlación.
r = -1	Existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante.

Fuente: elaboración propia, basada en la información de la página http://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson. Consulta julio de 2013.

En la tabla XLIII se muestra los resultados del cálculo de la correlación entre los factores de clima y el índice mismo de clima. De los cuales se puede concluir que todas las variables están altamente correlacionadas con el índice de clima organizacional. Sin embargo, es el factor de motivación el que tiene el índice más alto de correlación, con un valor de 0.94 que es muy cercano a una correlación perfecta.

Tabla XLIII. **Índices de correlación entre variables de clima**

Variable	Índice de Correlación r
Capacitación	0.90
Comunicación	0.92
Identificación e imagen	0.90
Liderazgo	0.89
Motivación	**0.94**
Organización laboral y cambio	0.91
Trabajo en equipo	0.85

Fuente: elaboración propia.

Concluido este análisis, se elige al factor motivación como la variable dependiente que será incluida en el modelo de regresión lineal múltiple.

2.6.3.2. Configuración de parámetros

El modelo que se busca encontrar tiene la siguiente forma:

$$Y' = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Dónde:

Y' = variable independiente (factor motivación)

a = intercepto en el eje Y

$X_1, X_2, \ldots X_n$ = variables independientes (liderazgo, capacitación, etc.)

$b_1, b_2, \ldots b_n$ = coeficientes de cambio de las variables independientes.

Los parámetros requeridos para generar el modelo de regresión son:

- Fórmula: es una descripción simbólica del modelo que debe ser encontrado. En el programa *R* se simboliza de la siguiente manera:

  variable_dependiente~variable_explicativa_1+variable_explicativa_2+…+

  En el caso específico de esta investigación este parámetro se configuró así:

  motivación ~ capacitación + comunicación + identificación + liderazgo + organización + trabajoequipo

  Es un modelo que trata de explicar la influencia que tienen las variables de clima con respecto a la variable motivación, como se observó en el apartado anterior es la variable que está más correlacionada con el índice de clima organizacional.

- Datos: el segundo parámetro que necesita el algoritmo es el conjunto de datos a partir del cual se realizarán los procesos para construir el modelo. En este caso este parámetro consta de 50 evaluaciones con datos del promedio de clima para cada una de los 7 factores seleccionados: motivación, capacitación, comunicación, identificación e imagen, liderazgo, organización laboral y trabajo en equipo. El formato de los datos es el que se presentó en la tabla XXXII.

2.6.3.3. Construcción del modelo

El proceso que se llevó a cabo para construir el modelo consistió en realizar llamadas a la función *lm* del software *R*, indicando los parámetros descritos en el numeral anterior. Esta función construye un modelo de regresión y proporciona el coeficiente del intercepto y los coeficientes de las

variables independientes, adicionalmente proporciona información para validar el modelo.

Debido a que existen 6 posibles variables explicativas, es necesario asegurarse que solo queden aquellas variables que expliquen de forma precisa el modelo. Inicialmente se incluyeron todas las variables y se aplicó proceso llamado eliminación progresiva que permite elegir el subconjunto de variables explicativas o independientes para el modelo. En concreto, la primera variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente o lo que es equivalente, si el p-valor asociado al estadístico T, o probabilidad de salida, es mayor que un determinado valor crítico, la variable es eliminada.[14]

El valor crítico utilizado en el proceso de eliminación progresiva es de 0.05.

En cada iteración en la que se elimina una variable, se observan los estadísticos de bondad de ajuste del nuevo modelo generado que se describen a continuación[15]:

- Suma de cuadrado de error: este estadístico mide la desviación total entre los valores de los datos y los de la regresión estimada. Un valor cercano a cero indica que el modelo tiene un componente aleatorio de error más pequeño y será más útil las predicciones que se hagan con él.

[14] GIBAJA, Juanjo. *Selección de variables explicativas en la regresión.* [en línea] <http://jjgibaja.net/seleccion-de-variables-explicativas-en-la-regresion/> [Consulta: 19 de julio de 2013].

[15] *Goodness of fit statistics.* [en línea] <http://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html> [Consulta: 10 de julio 2013].

- Coeficiente de determinación (R^2): mide que tan exitoso es el ajuste del modelo explicando la variación de los datos. Este estadístico toma valores entre 0 y 1, al tener un valor cercano a 1 indica que una mayor proporción de la varianza es predecida por el modelo.

- Coeficiente de determinación ajustado (ajuste de R^2): mide lo mismo que el coeficiente de determinación, con la diferencia de que este no está influenciado por el número de variables que introducimos. Generalmente este es el mejor indicador de la calidad de ajuste para comparar dos modelos. Puede tomar valores iguales o menores a 1, indicando 1 un mejor ajuste. Si hay números negativos el modelo quiere decir que el modelo contiene variables que no ayudan a predecir la respuesta**.**

A continuación se presentan los datos de las iteraciones ejecutadas para encontrar el modelo final. En la primera iteración se incluyen las 6 variables iniciales:

Motivación = Capacitación + Comunicación + Identificación + Liderazgo + Organización + Trabajo en equipo

En la tabla XLIV se muestran los estadísticos de bondad de ajuste para esta iteración.

Tabla XLIV. **Estadísticos de bondad de ajuste del modelo en la iteración 1**

Estadístico	Valor
Suma de cuadrados de error	0.07452
Coeficiente de determinación (R^2)	0.91450
Coeficiente de determinación ajustado	0.90260

Fuente: elaboración propia.

En la tabla XLV se muestra los coeficientes estimados, el error estándar, el valor T, el valor P y la significancia para cada variable. Acá es útil recordar que el valor P debe ser menor al valor crítico establecido de 0.05, para que la variable pueda ser considerada explicativa, de lo contrario deberá ser eliminada del modelo. En este caso las variables de comunicación y liderazgo no tienen significancia ya que el valor p es mayor que el valor crítico establecido. Sin embargo el proceso indica que solo debe ser eliminada una variable a la vez, por lo que se elegirá la que tiene el valor P más alto, que es la variable de comunicación.

Tabla XLV. **Coeficientes del modelo de la iteración 1**

Coeficientes	Estimado	Error Estándar	Valor T	Valor P	Significancia
Intercepto	-0.50431	0.21845	-2.309	0.02584	*
Capacitación	0.21948	0.08719	2.517	0.01564	*
Comunicación	-0.01148	0.12152	-0.094	0.92515	
Identificación	0.60265	0.08503	7.087	9.62E-09	***
Liderazgo	0.0905	0.08106	1.116	0.27043	
Organización	0.39072	0.12742	3.066	0.00374	**
Trabajo en equipo	-0.19036	0.09662	-1.97	0.05528	.

Fuente: elaboración propia.

En la segunda iteración se incluyen las 5 variables restantes:

Motivación = Capacitación + Identificación + Liderazgo + Organización + Trabajo en equipo

En la tabla XLVI se muestran los estadísticos de bondad de ajuste para esta iteración.

Tabla XLVI. **Estadísticos de bondad de ajuste del modelo en la iteración 2**

Estadístico	Valor
Suma de cuadrados de error	0.07367
Coeficiente de determinación (R^2)	0.91450
Coeficiente de determinación ajustado (ajuste de R^2)	0.90480

Fuente: elaboración propia.

De acuerdo a los valores P de la tabla XLVII se identifica a la variable de liderazgo como una variable que supera el valor crítico por lo cual deberá ser eliminada del modelo.

Tabla XLVII. **Coeficientes del modelo de la iteración 2**

Coeficientes	Estimado	Error Estándar	Valor T	Valor P	Significancia
Intercepto	-0.50483	0.21591	-2.338	0.02399	*
Capacitación	0.21579	0.0771	2.799	0.00758	**
Identificación	0.60298	0.084	7.178	6.27E-09	***
Liderazgo	0.08835	0.07691	1.149	0.25689	
Organización	0.38694	0.11961	3.235	0.00231	**
Trabajo en equipo	-0.19212	0.09373	-2.05	0.04639	*

Fuente: elaboración propia.

En la tercera iteración se incluyen las 4 variables restantes:

Motivación = Capacitación + Identificación + Organización + Trabajo en equipo

En la tabla XLVIII se muestran los estadísticos de bondad de ajuste para esta iteración.

Tabla XLVIII. **Estadísticos de bondad de ajuste en la iteración 3**

Estadístico	Valor
Suma de cuadrados de error	0.07393
Coeficiente de determinación (R^2)	0.9120
Coeficiente de determinación ajustado (ajuste de R^2)	0.9041

Fuente: elaboración propia.

De acuerdo a los valores P de la tabla XLIX se identifica a la variable de trabajo en equipo como una variable que supera el valor crítico por lo cual deberá ser eliminada del modelo.

Tabla XLIX. **Coeficientes del modelo de la iteración 3**

Coeficientes	Estimado	Error Estándar	Valor T	Valor P	Significancia
Intercepto	-0.51536	0.21648	-2.381	0.02158	*
Capacitación	0.24602	0.07273	3.383	0.00149	**
Identificación	0.59119	0.08367	7.066	8.13E-09	***
Organización	0.44618	0.1083	4.12	1.60E-04	***
Trabajo en equipo	-0.17875	0.09334	-1.915	0.06184	.

Fuente: elaboración propia.

En la cuarta iteración se incluyen las 3 variables restantes:

Motivación = Capacitación + Identificación + Organización

En la tabla L se muestran los estadísticos de bondad de ajuste para esta iteración.

Tabla L. **Estadísticos de bondad de ajuste del modelo en la iteración 4**

Estadístico	Valor
Suma de cuadrados de error	0.07605
Coeficiente de determinación (R^2)	0.9048
Coeficiente de determinación ajustado (ajuste de R^2)	0.8986

Fuente: elaboración propia.

De acuerdo a los valores P de la tabla LI se observa que las 3 variables restantes (capacitación, identificación y organización) tienen significancia para el modelo. Adicionalmente se observa que tanto el coeficiente de determinación y el coeficiente de determinación ajustado tienen valores cercanos a 1, por lo que indican un buen ajuste. También la suma de cuadrados de error tiene un valor cercano a cero, lo que indica que hay el componente de error aleatorio es muy pequeño.

Tabla LI. **Coeficientes del modelo de la iteración 4**

Coeficientes	Estimado	Error Estándar	Valor T	Valor P	Significancia
Intercepto	-0.6118	0.21656	-2.825	0.006969	**
Capacitación	0.23596	0.07461	3.163	0.002769	**
Identificación	0.56232	0.08465	6.643	3.13E-08	***
Organización	0.34343	0.09677	3.549	9.04E-04	***

Fuente: elaboración propia.

De acuerdo al análisis de esta última iteración se elige el modelo de la iteración 4 como modelo de regresión lineal múltiple definitivo para la explicación de la variable de motivación:

$$Motivación = -0.6118 + 0.23596\ Capacitación + 0.56232\ Identificación + 0.34343\ Organización$$

2.6.3.4. Interpretación de resultados

A partir de la construcción de los modelos de regresión lineal, en conjunto con el análisis de las encuestas de medición y las preguntas evaluadas se puede concluir lo siguiente.

La variable motivación y ambiente de trabajo es la que tiene una mayor correlación directa con el índice de clima organizacional, por lo cual resulta de mayor interés promover y generar planes de acción orientados a mejorar este aspecto del clima. Con base en los reactivos de las encuestas aplicadas por

Estratek, a continuación se enumeran algunos aspectos de la variable motivación y ambiente de trabajo que deben tomarse en cuenta:

- Los colaboradores esperan que la empresa premie el trabajo bien hecho.

- Los colaboradores valoran la relación de trabajo que se da entre compañeros.

- Los colaboradores necesitan contar con el equipo de trabajo adecuado para cumplir su trabajo.

Con respecto a la dependencia de la variable de motivación con respecto a las otras variables, se logró determinar teóricamente (utilizando el modelo de regresión lineal múltiple) que la motivación de los colaboradores está fuertemente influenciada por las siguientes 3 variables:

- Identificación e imagen: esta es la variable del modelo con un mayor coeficiente positivo, por lo cual se puede decir que es la que más impacta de las 3. Tomar en cuenta los siguientes aspectos puede ayudar a mejorar la calificación de esta variable:
  - Que la empresa muestre interés en el cuidado y bienestar de sus trabajadores.
  - Fomentar entre los colaboradores, la comunicación sobre los planes y propósitos de la empresa.

- Organización laboral y cambio: esta es la variable del modelo que le sigue a identificación e imagen en su influencia sobre la motivación de los colaboradores. Tomar en cuenta los siguientes aspectos puede ayudar a mejorar la calificación de esta variable:

- Asegurarse que los colaboradores conozcan bien todas las funciones que deben de realizar en su puesto de trabajo.
- Que el personal tenga claro cómo su desempeño individual contribuye al cumplimiento de la visión de la empresa.
- Comunicación adecuada sobre los cambios que se realizan en la empresa, cuáles son sus objetivos y el impacto que tuvieron.

- Capacitación y Desarrollo: esta es la variable que menos influye en el modelo. Sin embargo, si tiene significancia para haber formado parte del mismo. Algunos de los aspectos que se deben tomar en cuenta en la mejora de este factor son:
  - Ofrecer a los colaboradores oportunidades de capacitación.
  - Tomar en cuenta a los colaboradores de la empresa para promoverlos a plazas vacantes.
  - Que los colaboradores tengan claro cuáles son sus oportunidades de crecimiento y desarrollo dentro de la empresa.

2.6.3.5. Evaluar modelo

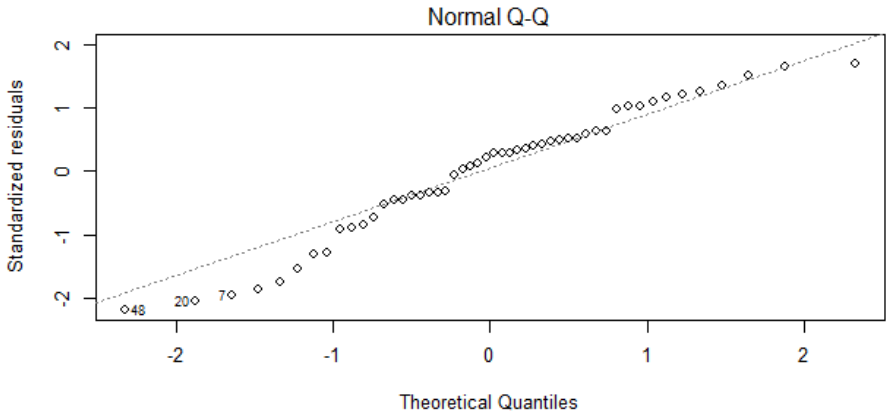
La validez del modelo de regresión lineal múltiple será evaluada desde varias perspectivas descritas a continuación.

2.6.3.5.1. Normalidad

La hipótesis de normalidad afirma que los errores del modelo siguen una distribución normal. Esta hipótesis se contrasta a partir de los residuos estandarizados. En la figura 9 de muestra un gráfico de probabilidad normal (que es aquel que indica que los datos tienen una distribución normal cuando los puntos caen sobre una recta) de los residuos estandarizados graficados en

función de las predicciones. Se puede observar que la mayor parte de datos caen sobre la recta, sin embargo, hay valores que no están alineados.

Figura 9. **Gráfica Q-Q de los residuos estandarizados**



Fuente: elaboración propia, generado con el programa informático *R.*

Debido a que no todos los puntos caen sobre la recta, se realizó una segunda validación por medio de la prueba de Shapiro-Wilk que permite comprobar la normalidad de un conjunto de datos. El programa *R* posee una función para aplicar este *test*. Se estableció un valor crítico de 0.05, si el valor-p devuelto por la prueba es mayor al valor crítico se concluye que los datos están distribuidos normalmente. Como se muestra en la tabla LII, el valor-p (0.07845) es mayor al valor crítico (0.05) por lo que se confirma la normalidad en los datos de los residuos estandarizados del modelo de regresión.

Tabla LII. **Resultado de la prueba de normalidad Shapiro-Wilk**

Estadístico	Valor
W	0.9587
Valor P	0.07845

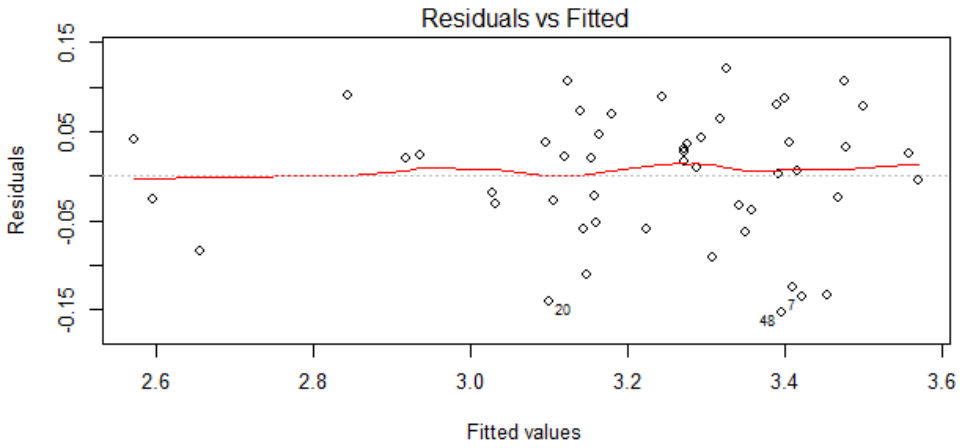
Fuente: elaboración propia.

2.6.3.5.2. Homocedasticidad

La homocedasticidad o igualdad de varianzas es una de las características que debe tener un modelo de regresión lineal. Una forma de validar si el modelo cumple con esta característica es graficar el conjunto de residuos versus el conjunto de las predicciones. Debido a que se supone que hay una igualdad de varianzas, el tamaño de los residuos es independiente del tamaño de los pronósticos, por lo cual en el diagrama de dispersión de residuos versus predicciones no debe notarse ningún patrón de asociación entre los 2 grupos.[16]

En la figura 10 se muestra la gráfica de dispersión entre residuos y predicciones del modelo de motivación y como se puede observar visualmente no se muestra ningún patrón o asociación entre los 2 grupos, por lo cual se puede concluir que el modelo tiene la propiedad de homocedasticidad.

Figura 10. **Gráfica de dispersión residuos versus predicciones**



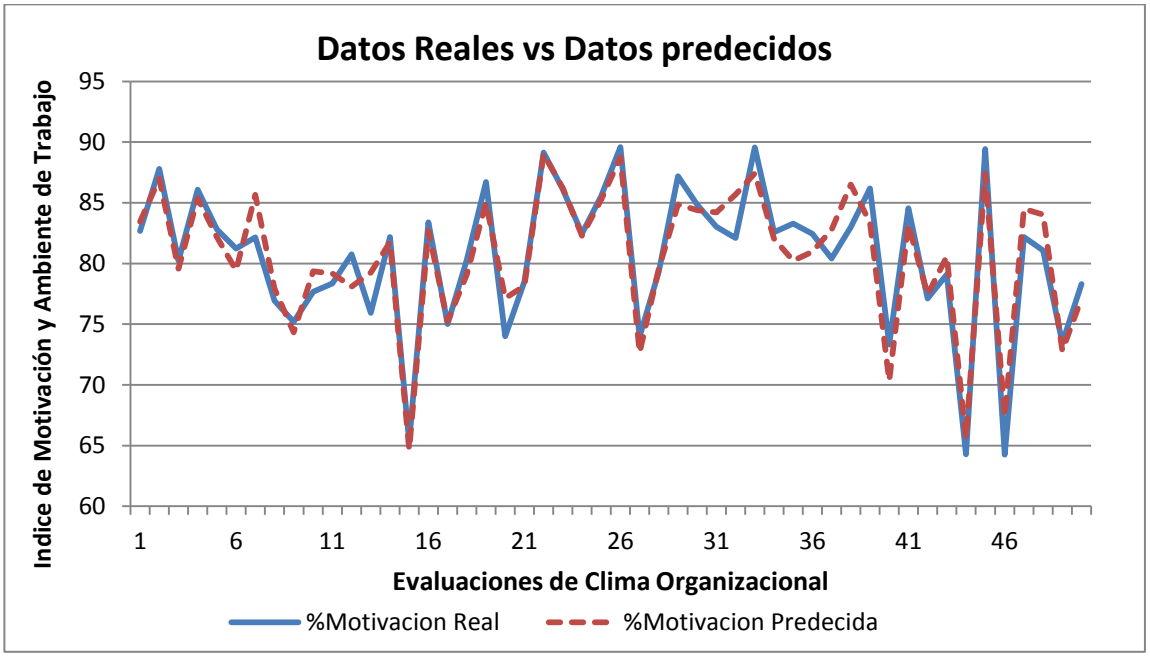
Fuente: elaboración propia, generado con el programa informático R.

[16] *Análisis de regresión lineal.* [en línea] <http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf> [Consulta: 20 de julio de 2013].

2.6.3.5.3. Predicción de resultados

Debido a que se cuenta con un modelo matemático fue posible realizar una predicción de datos para evaluar la precisión del mismo. En la figura 11 se muestra una gráfica de líneas dónde se trazan los valores reales versus las predicciones y se puede observar que existe un buen nivel de precisión en las predicciones hechas por el modelo.

Figura 11. **Gráfica de predicción del modelo versus datos reales**



Fuente: elaboración propia.

CONCLUSIONES

1. En las investigaciones aplicadas al estudio de las ciencias humanas, el desarrollo de modelos matemáticos, estadísticos o de información, permiten comprobar si las inferencias realizadas son consistentes con los datos analizados. Sin embargo, esta consistencia entre los datos y el modelo no implica obligatoriamente una consistencia entre el modelo y la realidad. Únicamente se puede afirmar que los modelos e inferencias propuestas no son contradictorios y por ende estos pueden ser válidos.

2. Aplicar la metodología estándar CRISP-DM en el proceso de desarrollo de la minería de datos fue una excelente decisión, porque permitió tener claro las fases necesarias y un conjunto de mejores prácticas para el desarrollo del proyecto.

3. La técnica de *clustering* desde el punto de vista de minería de datos, es uno de los modelos básicos que generalmente se implementa como un primer paso, para la solución de otro tipo de problemas más complejos. Su propósito principal es mantener el tamaño de los datos manejables y encontrar subconjuntos de datos que sean más fáciles de analizar.

4. La técnica de *clustering* utilizada permitió segmentar la información de evaluaciones de clima organizacional en 3 grupos. Cada grupo representado por sus características demográficas y por el nivel de satisfacción de sus colaboradores: alto, promedio y bajo. Al aplicar este conocimiento en temas como los procesos de selección de personal para

puestos críticos, puede tener implicaciones positivas para la contratación de colaboradores con alto compromiso y motivación.

5. La construcción del modelo de regresión lineal múltiple, para explicar la motivación de los colaboradores de una empresa a través de otras variables de clima organizacional, permitió descubrir que la motivación es una variable dependiente de: 1) la identificación e imagen que tiene el colaborador con la empresa, 2) la organización laboral y la adaptación al cambio y 3) la capacitación y desarrollo que percibe el colaborador dentro de la empresa.

RECOMENDACIONES

1. Antes de iniciar un proceso de minería de datos, es indispensable realizar una planificación detallada del mismo por lo que se recomienda utilizar una metodología estándar que permita establecer los objetivos, alcances, recursos, riesgos y proceso de implementación. Realizarlo de esta manera garantizará en un alto porcentaje, el éxito del proyecto.

2. En todo proyecto de minería de datos es indispensable la validación de los modelos de minería de datos, ya que permite tener la certeza de que el conocimiento descubierto tiene validez y puede convertirse en un insumo de calidad para la toma de decisiones.

3. Las instituciones interesadas en realizar minería de datos, deben realizar un esfuerzo en depurar los datos que administran y que sirven de insumo para las técnicas de minería. Uno de los principales inconvenientes que se presentan en este tipo de proyectos, son los datos incompletos o incorrectos los cuales deben ser descartados y que hubieran podido aportar información valiosa para la construcción de los modelos.

4. Se sugiere utilizar los resultados de este proyecto como base para futuros trabajos, los cuales pueden incorporar técnicas de minería de datos más complejas como la clasificación, árboles de decisión, etc. Así como la incorporación de otras variables de gestión empresarial como: nivel de desempeño, rentabilidad de la empresa, rotación de colaboradores, etc. La evolución de este tipo de análisis, permite encontrar otro tipo de conocimiento que permita tomar acciones en busca

de la mejora de condiciones laborales, cultura organizacional y otros temas de la gestión del talento humano.

BIBLIOGRAFÍA

1. ALONSO LLOMBART, Oscar. *Metodología CRISP-DM.* [en línea] <http://es.slideshare.net/oalonso/metodologa-de-data-mining-crisp> [Consulta: 22 de junio de 2013].

2. CALLEJA PALMA, Carmen. *La Evaluación del clima organizacional como alternativa de mejora de la productividad laboral.* [en línea] <http://www.bibliotecadigital.uson.mx/bdg_tesisIndice.aspx?tesis=21317> Trabajo de graduación, Universidad de Sonora, México. [Consulta: 19 de abril de 2013].

3. *¿Cómo hacer una encuesta?* [en línea] <http://www.rrppnet.com.ar/comohacerunaencuesta.htm> [Consulta: 15 de mayo de 2013].

4. *Conceptos de minería de datos.* [en línea] <http://msdn.microsoft.com/es-es/library/ms174949.aspx> [Consulta: 05 de junio de 2013].

5. *Cross Industry Standard Process for Data Mining.* [en línea] <http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining> [Consulta: 25 de junio de 2013].

6. *Data mining algorithms in R.* [en línea] <http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R> [Consulta: 20 de mayo de 2013].

7. DISDIER, Orville. *Introducción a R, programa estadístico de "código abierto".* [en línea] <http://es.slideshare.net/ODISDIER/introduccion-al-programa-r> [Consulta: 27 de mayo de 2013].

8. *Documentación del software SciKit learning.* [en línea] <http://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods> [Consulta: 7 de junio de 2013].

9. *Escalas Likert.* [en línea] <http://es.wikipedia.org/wiki/Escalas_Likert> [Consulta: 20 de mayo de 2013].

10. GARCÍA SANTIESTEBAN, David. *Minería de datos para la detección de patrones criminalísticos en Guatemala*. Trabajo de graduación de Ing. en Ciencias y Sistemas. Universidad de San Carlos de Guatemala, Facultad de Ingeniería, 2012. 169 p.

11. GOICOCHEA, Aníbal. *CRISP-DM, Una metodología para proyectos de Minería de Datos*. [en línea] <http://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/> [Consulta: 29 de junio de 2013].

12. GONCALVES, Ángel P. *El Clima como término organizacional.* [en línea] <http://moodle.unid.edu.mx/dts_cursos_mdl/maestria_en_educacion/desarrollo_y_com_en_los_r_h/sesion4/actividades/ClimaTerminoOrganizacional.pdf> [Consulta: 9 de mayo de 2013].

13. IBM Corporation. *IBM SPSS modeler CRISP-DM guide*. [en línea] <ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf> [Consulta: 19 de junio de 2013].

14. *Instrumentos para medir el clima organizacional.* [en línea] <http://www.slideboom.com/presentations/220161/Instrumentos-para-medir-el-clima-organizacional> [Consulta: 19 de mayo de 2013].

15. LORENZ, Honrad. *Escala de clima organizacional* (EDCO). [en línea] <http://www.monografias.com/trabajos35/escala-clima-organizacional/escala-clima-organizacional.shtml> [Consulta: 03 de junio de 2013].

16. *Minería de datos*. [en línea] <http://campusvirtual.unex.es/cala/epistemowikia/index.php?title=Miner%C3%ADa_de_Datos#Tipos_de_Modelos_de_Datos> [Consulta: 09 de junio de 2013].

17. MOLINA, Maisch. *Estudios de clima organizacional.* [en línea] <http://www.losrecursoshumanos.com/contenidos/290-estudios-de-clima-organizacional.html> [Consulta: 7 de mayo de 2013].

18. *¿Qué es una encuesta?* [en línea] <http://www.portaldeencuestas.com/que-es-una-encuesta.php> [Consulta: 10 de junio de 2013].

19. SULLIVAN, Dan. *The cloud: how to validate data mining models.* [en línea] <http://www.tomsitpro.com/articles/data_mining_as_a_service-method_of_testing-classification_algorithm,1-256.html> [Consulta: 08 de julio de 2013].

20. TAN, Steinbach. *Data mining cluster analysis: basic concepts and algorithms.* [en línea] <http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap8_basic_cluster_analysis.pdf> [Consulta: 09 de junio de 2013].

21. *Testing and validation (Data Mining)*. [en línea] <http://technet.microsoft.com/en-us/library/ms174493.aspx> [Consulta: 05 de julio de 2013].

22. *What are the different types of data mining analysis?* [en línea] <http://www.wisegeek.com/what-are-the-different-types-of-data-mining-analysis.htm> [Consulta: 19 de junio de 2013].

23. ZAHII, Mohammed; MEIRA, Wagner. *Data mining and analysis: foundations and algorithms.* [en línea] <http://www.dcc.ufmg.br/miningalgorithms/files/pdf/fdma.pdf> [Consulta: 22 de junio de 2013].

ANEXOS

CONTRATO DE CONFIDENCIALIDAD PARA EL USO DE LA INFORMACIÓN DE ESTRATEK

CONTRATO DE CONFIDENCIALIDAD PARA EL DESARROLLO DE LA TESIS "MINERÍA DE DATOS PARA ENCONTRAR PERFILES DE COLABORADORES SATISFECHOS Y LOS FACTORES QUE INCIDEN EN SU SATISFACCIÓN", QUE CELEBRAN POR UNA PARTE PABLO FERNANDO MENDOZA ZEPEDA, EN LO SUCESIVO DENOMINADO "EL INVESTIGADOR"; POR LA OTRA PARTE, LUIS ALBERTO GARCÍA RUANO, PROPIETARIO DE LA EMPRESA ESTRATEK EN LO SUCESIVO DENOMINADO "ESTRATEK", EL CUAL SE SUJETA A LAS SIGUIENTES DECLARACIONES Y CLAUSULAS:

PROPIEDAD.

La licencia de "Clima Organizacional ESTRATEK" es y seguirá siendo en todo momento, propiedad del titular "ESTRATEK". "EL INVESTIGADOR" no tendrá ningún derecho, titulo o interés en el programa de "CLIMA ORGANIZACIONAL ESTRATEK", y no permitirá que ninguna obligación o gravamen exista sobre éste, ni permitirá el uso de "CLIMA ORGANIZACIONAL ESTRATEK" por terceros, ni realizará ningún acto que pueda modificar los derechos de autor de "CLIMA ORGANIZACIONAL ESTRATEK". Por lo tanto, "EL INVESTIGADOR" no puede vender, arrendar, prestar, revelar, transmitir o transferir a ningún título, ni copiar el programa de "CLIMA ORGANIZACIONAL ESTRATEK" para terceros.

CONFIDENCIALIDAD.

La licencia y el programa de "CLIMA ORGANIZACIONAL ESTRATEK", como también cualquier documentación e información relacionada con éste, son propiedad exclusiva del titular "ESTRATEK". "EL INVESTIGADOR" conviene en mantener en estricta confidencialidad toda información de carácter confidencial, ya que toda la mencionada información es propiedad del titular "ESTRATEK".

La información que "ESTRATEK" proporcione a "EL INVESTIGADOR" para el desarrollo de la investigación objeto de este contrato, así como la que se desprenda con motivo y como consecuencia del mismo, será considerada como estrictamente confidencial. "EL INVESTIGADOR" se obliga a no revelar dicha información a terceros, ajenos a las partes, y a utilizarla exclusivamente para la ejecución de este contrato. "EL INVESTIGADOR" garantiza a "ESTRATEK" que tomarán las providencias precautorias que sean necesarias para evitar que la personas que lleguen a manejar la información propiedad de "ESTRATEK" no la divulguen y que terceros ajenos a las partes no tengan acceso a ella.

Ambas partes se comprometen a tratar de manera confidencial cualquier información técnica o científica que se proporcione y a la que tuvieren acceso a partir de la vigencia de este contrato.

"EL INVESTIGADOR", se compromete por este acto, a cumplir estrictamente las instrucciones y procedimientos establecidos por "ESTRATEK" para el uso adecuado de "CLIMA ORGANIZACIONAL ESTRATEK" y a su vez, se compromete a instruir a sus dependientes acerca de estos procedimientos e instrucciones.

Dentro de esta información se contemplan proyectos, compilaciones, estudios, datos informativos y toda aquella documentación elaborada o proporcionada por las partes o compañías que se relacionen con ellas y sus representantes o bien sus socios, trabajadores, directores, empleados, consultores, contadores, auditores; también contempla toda aquella información a la cual tuviere acceso durante la prestación de los servicios contratados que, al momento de ser divulgada a la competencia, tuviere algún impacto relevante.

El término información, no incluye la información que pueda ser obtenida por cualquier persona o aquella que sea generalmente proporcionada públicamente.

Obligaciones: Bajo el acuerdo de confidencialidad, serán obligaciones de las partes:

a)Usar la información que se le proporcione únicamente para llevar a cabo el desempeño de los servicios acordados; b) mantener en estricta confidencialidad, la información que se le proporcione, información que no deberá ser utilizada en detrimento de los intereses de las partes; c) notificar a la otra parte inmediatamente de cualquier petición efectuada por autoridad competente para la revelación de la información, acorde con lo establecido en el presente acuerdo.

Incumplimiento: En caso de que las partes incumplieren con algunas de las disposiciones contenidas en el presente ACUERDO DE CONFIDENCIALIDAD, será responsable de los daños y perjuicios que causare a la otra parte por la divulgación de la información relacionada.

Plazo: Los derechos y obligaciones contemplados en el presente acuerdo serán aplicables por un plazo indefinido.

ACEPTACIÓN GENERAL.

Tanto ESTRATEK a través de su representante legal como PABLO FERNANDO MENDOZA ZEPEDA, manifestamos ACEPTACIÓN EXPRESA a lo pactado en este contrato, y enterados de su contenido, validez y efectos legales, lo ratificamos y firmamos.

"EL INVESTIGADOR	"ESTRATEK"
_____	_____
Pablo Fernando Mendoza Zepeda	Luis Alberto García Ruano
Cédula: A-1 1099085	Cédula: A-1 1100193
Investigador	Representante Legal

.