UNIVERSIDAD DE SONORA
DIVISIÓN DE INGENIERÍA

POSGRADO EN INGENIERÍA INDUSTRIAL
MAESTRÍA EN INGENIERÍA EN SISTEMAS Y TECNOLOGÍA

OPTIMIZACIÓN DEL PROCESO DE RECLUTAMIENTO EN UNA EMPRESA DE DESARROLLO DE SOFTWARE CON TÉCNICAS DE MINERÍA DE DATOS

T E S I S

PRESENTADA POR

JESUS ABIRAN LOPEZ RAMIREZ

Desarrollada para cumplir con uno de los
requerimientos parciales para obtener
el grado de Maestro en Ingeniería

DIRECTORA DE TESIS
DRA. RAQUEL TORRES PERALTA

CODIRECTOR
DR. MARIO BARCELÓ VALENZUELA

HERMOSILLO, SONORA, MÉXICO. SEPTIEMBRE 2017

"El saber de mis hijos hará mi grandeza"

Hermosillo, Sonora a 10 de agosto de 2017

JESUS ABIRAN LOPEZ RAMIREZ

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado vigente, otorgamos a usted nuestra aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestro en Ingeniería.

Por tal motivo este jurado extiende su autorización para que se proceda a la impresión final del documento de tesis: **OPTIMIZACIÓN DEL PROCESO DE RECLUTAMIENTO EN UNA EMPRESA DE DESARROLLO DE SOFTWARE CON TÉCNICAS DE INTELIGENCIA ARTIFICIAL Y MINERÍA DE DATOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

Dra. Raquel Torres Peralta
Directora de Tesis y Presidente del Jurado

Dr. Mario Barceló Valenzuela
Codirector y Vocal del Jurado

Dr. Federico Miguel Cirett Galán
Secretario del Jurado

Dr. Alonso Pérez Soltero
Vocal del Jurado

c.c.p. Archivo

UNIVERSIDAD JUÁREZ
AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DAIS
11111000011

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

Villahermosa, Tabasco, México, a 29 de junio de 2017.

JESUS ABIRAN LOPEZ RAMIREZ

Con fundamento en el articulo 66, fracción III, del Reglamento de Estudios de Posgrado de la Universidad de Sonora, otorgo a usted mi aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestro en Ingeniería.

Por tal motivo, como sinodal externo y vocal del jurado, extiendo mi autorización para que se proceda a la impresión final del documento de tesis: **OPTIMIZACIÓN DEL PROCESO DE RECLUTAMIENTO EN UNA EMPRESA DE DESARROLLO DE SOFTWARE CON TÉCNICAS DE MINERÍA DE DATOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

DR. PABLO PAYRÓ CAMPOS
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
Sinodal Externo y Vocal del Jurado

Miembro CUMEX desde 2008
Consorcio de Universidades Mexicanas

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690, Cunduacán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0516; Fax: (914) 336 0870

RESUMEN

En la actualidad se reconoce el impacto que tiene en las organizaciones el que se contrate al personal mas capacitado y con mejores habilidades para desarrollar las funciones dentro de la misma, es por eso que es importante prestar especial atención al proceso de reclutamiento para lograr este objetivo.

La minería de datos es una rama de la inteligencia artificial que hoy en día tiene muchas aplicaciones. Se utiliza principalmente para predicciones fiables de sucesos con base en datos históricos. Esta investigación presenta una propuesta para optimizar los procesos de selección de personal en una empresa desarrolladora de software, identificando a los candidatos con mayor probabilidad de tener éxito en las distintas etapas del proceso de selección de personal con el fin de acelerar el proceso de evaluación.

Mi enfoque evalúa el desempeño de diferentes algoritmos de aprendizaje de máquinas supervisado para predecir el desempeño de los candidatos durante el proceso de selección utilizando como atributos principales las respuestas a evaluaciones técnicas, tomando en cuenta sólo sus aptitudes técnicas y de resolución de problemas. Los resultados en experimentos con varios algoritmos de clasificación arrojan hasta un 100% de acierto en las predicciones.

Este tipo de enfoque permite detectar con anticipación a los mejores candidatos, pero además excluye la información sociodemográfica que pudiera influir en la decisión final, evitando un sesgo discriminatorio.

Los principales resultados obtenidos de la implementación del modelo, fueron la identificación de candidatos con mayor probabilidad de ser contratados desde una etapa temprana dentro del largo proceso de reclutamiento de esta empresa, lo cual es de gran utilidad para acelerar su proceso de contratación.

ABSTRACT

At present, it is recognized the impact that has on the organizations the recruitment of the most qualified personnel with the best skills to develop the functions within the organization, this is why it is important to pay special attention to the recruitment process to achieve this goal.

Data mining is a branch of artificial intelligence that today has many applications. It is mainly used for reliable predictions of events based on historical data. This research presents a proposal to optimize the personnel selection processes in a software development company, identifying the candidates most likely to be successful in the different stages of the personnel selection process in order to accelerate the evaluation process.

My approach evaluates the performance of different supervised machine learning algorithms to predict candidate performance during the selection process using as main attributes responses to technical assessments, considering only their technical and problem-solving skills. The results in experiments with several classification algorithms show up to 100% accuracy in the predictions.

This type of approach allows early detection of the best candidates, but also excludes sociodemographic information that could influence the final decision, avoiding a discriminatory bias.

The main results obtained from the implementation of the model were the identification of candidates with higher probability of being hired from an early stage within the long recruitment process of this company, which is very useful to accelerate its process of evaluation.

DEDICATORIAS

A mi esposa Polet. Tu ayuda ha sido fundamental, has estado conmigo incluso en los momentos más turbulentos. Este proyecto no fue fácil, pero estuviste motivándome y ayudándome hasta donde tus alcances lo permitían. Por esto, ahora puedo decir que soy mejor persona. Te amo.

A mis padres Jesús y Marina por haberme forjado como la persona que soy en la actualidad; muchos de mis logros se los debo a ustedes entre los que se incluye este. Me formaron y motivaron constantemente para alcanzar mis anhelos.

A mis hermanos Dulce y Luis por ser una piedra fundamental en mi vida.

AGRADECIMIENTOS

Primeramente a Dios, por permitirme alcanzar una nueva meta en mi vida.

A mi esposa Polet por tu apoyo incondicional durante la estancia de mi posgrado. No fue sencillo culminar este proyecto pero siempre creiste en mí y nunca me soltaste de tu mano.

Gracias a mis padres por ser los principales promotores de mis sueños. A mi papá por confiar y creen en mí y en mis expectativas, a mi mamá porque durante el tiempo que Dios le prestó vida, me enseñó a salir adelante y forjó mi carácter.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa de Fortalecimiento de la Calidad Educativa (PFCE) por las becas otorgadas y los apoyos económicos brindados para contar con los recursos para realizar mis estudios de posgrado.

Al Departamento de Ingeniería Industrial de la Universidad de Sonora por permitirme cursar mis estudios de maestría.

A la Dra. Raquel Torres Peralta por ser mi directora y brindarme su apoyo incondicional, darme palabras de aliento y tener paciencia y confianza en mi, ya que esto hizo que pudiera concluir mi investigación. ¡Infinitas gracias!

A los Dres. Alonso Peréz Soltero y Mario Barcelo Valenzuela por sus revisiones y críticas constructivas hacia mi investigación.

A Nearsoft por darme la oportunidad de realizar mi investigación en esa empresa.

A Isaac López por ayudarme con la investigación dentro de la empresa.

A mis compañeros de posgrado, por cientos de horas que pasamos juntos y habernos convertido en complices en momentos tan importantes como este. Especialmente a Gerardo Sánchez e Ismael Camarena.

A mis maestros de posgrado por su enseñanza y guia.

A mi amigo Misael Gómez por motivarme a cursar y terminar este posgrado.

Finalmente quiero agradecer a todas aquellas personas que de una u otra manera me ayudaron durante mi estancia en el posgrado y durante la elaboración de esta tesis. A todos gracias.

A todos con mucho cariño.

ÍNDICE GENERAL

RESUMEN i
ABSTRACT ii
DEDICATORIAS iii
AGRADECIMIENTOS iv
ÍNDICE GENERAL vi
ÍNDICE DE FIGURAS viii
ÍNDICE DE TABLAS ix
1. INTRODUCCIÓN 1
1.1. Presentación 3
1.2. Planteamiento del problema 5
1.3. Objetivo general 5
1.4. Objetivos específicos 5
1.5. Hipótesis 6
1.6. Alcances y delimitaciones 6
1.7. Justificación 6
2. MARCO DE REFERENCIA 7
2.1. Recursos Humanos 7
2.2. Aprendizaje Automatizado 8
2.3. Minería de Datos 8
2.3.1. Clasificación y Predicción 10
2.3.2. Adquisición de Datos 11
2.3.3. Preparación de los Datos 11
2.4. Minería de Datos como herramienta en Recursos Humanos 12
2.5. Trabajos previos 13
2.6. Usos más comunes 14
2.7. Casos de éxito 19
2.8. ¿Por qué no es suficiente o qué vamos a mejorar? 22
2.9. Comparativa de los algoritmos para Minería de Datos 22
2.9.1. Naïve Bayes 23
2.9.2. Máquinas de Vectores de Soporte 24
3. MODELO 26
3.1. Definición de la problemática y estructuración del objetivo. 27
3.1.1. Definir el problema y las restricciones 28
3.1.2. Definir el objetivo de la minería de datos 28
3.1.3. Definir los atributos y variables 28
3.1.4. Seleccionar técnicas de minería de datos 29
3.2. Recolección y preparación de los datos. 29

3.2.1. Recopilación de los datos 30
3.2.2. Revisar la distribución de los datos 30
3.2.3. Procesamiento de datos 31
3.3. Construcción del modelo de minería de datos 34
3.3.1. Implementación de técnicas de minería de datos 34
3.4. Análisis de evaluación del modelo. 35
3.4.1. Someter el modelo a pruebas 36
3.5. Implementación y validación. 37
3.5.1. Implementación del modelo resultante de minería de datos 37
3.5.2. Medición de los resultados 37
3.5.3. Evaluación de los resultados 37
4. IMPLEMENTACIÓN 39
4.1. Definición de la problemática y estructuración del objetivo. 39
4.1.1. Definir el problema y las restricciones 39
4.1.2. Definir el objetivo de la minería de datos 40
4.1.3. Definir los atributos y variables 40
4.1.4. Seleccionar técnicas de minería de datos 41
4.2. Recolección y preparación de los datos. 41
4.2.1. Recopilación de los datos 41
4.2.2. Revisar la distribución de los datos 41
4.2.3. Procesamiento de datos 42
4.3. Construcción del modelo de minería de datos 44
4.3.1. Implementación de técnicas de minería de datos 44
4.4. Análisis de evaluación del modelo. 45
4.4.1. Someter el modelo a pruebas 45
4.5. Implementación y validación. 48
4.5.1. Implementación del modelo resultante de minería de datos 48
4.5.2. Medición de los resultados 48
4.5.3. Evaluación de los resultados 49
4.6. Ventajas de la Minería de Datos sobre la Heurística 51
5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS 55
5.1. Conclusiones 55
5.2. Recomendaciones. 56
5.3. Trabajos Futuros. 57
6. REFERENCIAS 58

ÍNDICE DE FIGURAS

Figura 2.1 Etapas de la Minería de Datos 9

Figura 3.1 Modelo para minería de datos. 27

Figura 3.2 Ciclo de procesamiento de datos. 32

Figura 4.1 Estructura de la tabla de registros conteniendo las respuestas al examen técnico y los resultados de las etapas 1 a la 4, que fueron utilizados para el entrenamiento del modelo de predicción. 43

ÍNDICE DE TABLAS

Tabla 4.1 Evaluación de etapa 1. 50

Tabla 4.2 Evaluación de etapa 2. 50

Tabla 4.3 Frecuencias de respuestas en el éxamen de lógica 52

Tabla 4.4 Relacion entre respuesta correcta y respuesta con mayor incidencia en el examen lógico 53

1. INTRODUCCIÓN

La selección de recursos humanos calificados es un factor clave de éxito para una organización, especialmente debido al aumento de los mercados competitivos y la globalización. La selección de personal es el proceso de elección de los individuos (candidatos) que satisfacen las calificaciones requeridas para realizar un trabajo definido de la mejor manera posible y tiene por objeto elegir el mejor candidato para llenar una vacante específica en una empresa u organización; el proceso de selección de personal tiene un impacto directo en la calidad del personal y por lo tanto juega un papel importante en la gestión de los recursos humanos y el éxito futuro de la empresa depende en gran medida de la contribución de su personal (Bello *et al.*, 2016).

El capital humano es una de las competencias básicas para que las empresas de tecnología mantengan su ventaja competitiva en la economía del conocimiento. El nivel de los empleados se ve afectado por el proceso de reclutamiento de las empresas. Se han realizado estudios sobre currículum vitae, entrevistas, centros de evaluación, pruebas de trabajo, cognitivas y de personalidad con el fin de ayudar a las organizaciones a tomar mejores decisiones para la selección de personal (Chien y Chen, 2008).

Debido a las diferentes tecnologías en la industria del desarrollo de software, los perfiles de trabajo pueden no ser tan fácilmente delimitables, especialmente para los puestos de trabajo de Desarrollo de Software (SD). A medida que las tecnologías avanzan, los nuevos perfiles de empleado requieren mayor conocimiento sobre las mismas. Los puestos de trabajo requieren personal con más y diversa experiencia por lo que el proceso de selección se hace más complicado. Por lo tanto, los enfoques convencionales de selección de personal que se desarrollan sobre la base de las características del trabajo estático ya no serán suficientes (Lievens *et al.*, 2002). Con el fin de encontrar a las personas correctas para hacer un buen trabajo, es vital desarrollar enfoques de selección efectivos.

La selección de personal juega un rol decisivo en la gestión de recursos humanos para determinar la calidad del personal contratado. Investigadores como Robertson y Smith, (2001) y Borman *et al.*, (1997), han revisado estudios de selección de personal y encontraron que situaciones como cambio en organizaciones, roles, normas de trabajo han influenciado en la selección de personal. Mientras que los avances en las tecnologías de información también afectan a la selección de personal en la gestión de recursos humanos (Kovach y Cathcart, 1999; Liao, 2003; Beckers y Bsat, 2002). La aplicación de sistemas de expertos o sistemas para ayuda en la toma de decisiones en reclutamiento y selección de personal ha crecido (Hooper *et al.*, 1998; Nussbaum *et al.*, 1999).

Recientemente, debido a los avances en las tecnologías de información, los investigadores han desarrollado sistemas para la ayuda en la toma de decisiones y sistemas expertos para mejorar los resultados de la gestión de recursos humanos. En particular, la minería de datos es reconocida como uno de los temas más sobresalientes. La minería de datos se refiere a la extracción o uso de patrones de una gran base de datos a través de una exploración y análisis automático o semiautomático (Chen *et al.*, 1996; Berry y Linoff, 1997). Con la ayuda de las técnicas de minería de datos, las computadoras ya no se limitan a solo guardar datos sino que pueden ayudar a los usuarios a extraer puntos clave de grandes cantidades de datos para usarlos como análisis o predicción.

Las metodologías de minería de datos se han desarrollado para la exploración y el análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos para descubrir patrones y reglas significativas. Este tipo de datos, incluyendo los datos del personal pueden proporcionar un rico recurso para el descubrimiento de conocimiento y apoyo a la decisión. La minería de datos es guiada por descubrimiento, no por supuestos. La minería de datos implica diversas técnicas que incluyen estadísticas, redes neuronales, árboles de decisión, algoritmos genéticos, y las técnicas de visualización que se han desarrollado en los últimos años.

La minería de datos se ha aplicado en muchos campos como la mercadotecnia, finanzas, banca, industria manufacturera, atención sanitaria, gestión de relaciones con clientes, detección de fallos y predicción, y la organización de aprendizaje (Chien *et al.*, 2002; Chien *et al.*, 2004; Shaw et al., 2001; Wei y Chiu, 2002; Wu *et al.*, 2005; Peng y Chien, 2003; Shiue y Su, 2003).

1.1. Presentación

La investigación se desarrolló en Nearsoft, una empresa mexicana de outsourcing que se dedica al desarrollo de software. La empresa cuenta con 3 oficinas en México localizadas en las ciudades de Hermosillo, Ciudad de México y Chihuahua.

Dentro de Nearsoft existen diversos departamentos, tales como Gerencia, Finanzas, Reclutamiento y Desarrollo Personal; dentro del área operativa se encuentran los departamentos de Diseño Gráfico, Desarrollo de Software (SD) y Aseguramiento de la calidad (QA), siendo los departamentos del área operativa los que generan los principales ingresos para la empresa puesto que ellos producen el servicio que se vende.

La empresa presta sus servicios a compañías internacionales donde los estándares de evaluación son elevados, de modo que la calidad de sus empleados se refleja directamente en los resultados esperados por el cliente. Las restricciones de tiempo en las entregas y en la calidad de las soluciones desarrolladas son producto de la preparación y capacidad de los programadores. La elección de candidatos es crucial para alcanzar los objetivos, mantener la relación con el cliente y cuidar el prestigio de la empresa para el crecimiento del negocio.

Dentro del mercado laboral, Nearsoft ofrece una gran cantidad de vacantes para diferentes puestos, por ejemplo para 2016 se necesitaban cubrir 80 vacantes, la mayoría de ellas fueron para el departamento de SD.

El proceso de reclutamiento es complejo, puesto que consta de 7 etapas por las cuales cada uno de los candidatos debe de pasar, las cuales son: Primer contacto(1), examen de lógica(2), entrevista en inglés(3), entrevista técnico

ligera(4), entrevista técnico profunda(5), entrevista cultural(6) y programación conjunta(7).

Cada una de las etapas de reclutamiento se califica en base a una escala numérica y cada una de las calificaciones va acompañada con una breve justificación del por qué las personas que estuvieron involucradas en esa parte de la entrevista asignaron ese valor; regularmente es un promedio de 6 personas diferentes incluyendo parte del equipo al cual el candidato aspira y el reclutador. Cada una las de fases del proceso de reclutamiento requiere una inversión de tiempo de parte de los involucrados lo que representa logística del equipo de reclutamiento y esfuerzo de parte de los entrevistados. El promedio de tiempo invertido a un candidato es 8 horas, desde el primer contacto hasta que acepta la oferta. Todo este tiempo se extiende a lo largo de 5 semanas en promedio.

Si bien la habilidad técnica es un requerimiento indispensable al momento de buscar a un candidato, la empresa busca de igual manera de cada uno de sus integrantes tenga en común una lista de valores que son inclusive más importantes que la habilidad técnica.

La lista de valores de la empresa es:

1. Compromiso: No importa lo que suceda, los miembros entregan su trabajo con la calidad esperada.
2. Liderazgo: Es importante que las personas demuestren un liderazgo natural independientemente la posición en la que se encuentren.
3. Relaciones a largo plazo: Se debe de poner todo el esfuerzo posible en mantener una buena relación a largo plazo con los compañeros, cliente y equipo.
4. Ser inteligente y hacer que las cosas sucedan: Ser inteligente no es suficiente, es importante tener la iniciativa y lograr que las cosas sucedan.
5. Trabajo en equipo: Saber trabajar en equipo con buena comunicación y relación entre los miembros para producir resultados de la mejor calidad.

Nearsoft siempre busca conseguir a los mejores profesionales para llenar sus vacantes. En 2015 requería cubrir 60 vacantes y sólo el 0.95% de los candidatos fueron contratados. La principal razón es que los postulantes no cumplían con el perfil. Existen candidatos con gran potencial que aprueban cada una de las fases del proceso de reclutamiento y al final rechazan una oferta económica dado a que el proceso es extenso y al tener mucha experiencia, en el intermedio reciben ofertas de otras empresas las cuales aceptan.

Hasta el día de hoy se tiene una base de datos de aproximadamente 4,100 personas las cuales han estado en el proceso de reclutamiento, ya sea con éxito o no, dentro de las cuales el 83% son hombres y 17% mujeres. Esta información muestra el puntaje obtenido en cada candidato a lo largo de su proceso, el cual puede servir para identificar patrones de comportamiento y de esta manera poder predecir si un candidato va a tener éxito a lo largo de su proceso.

1.2. Planteamiento del problema

Nearsoft no tiene una forma de saber cuáles de los candidatos van a pasar las etapas de entrevista por lo cual no puede priorizar sus procesos de reclutamiento y ocasiona que muchas veces pierda a los postulantes con mayor experiencia. Una gran cantidad de candidatos contactan o son contactados por la empresa para pasar por un largo proceso de reclutamiento. Este proceso es largo y exhaustivo para asegurar la calidad en sus empleados.

1.3. Objetivo general

Proponer y evaluar un algoritmo de minería de datos cuyo procedimiento sea capaz de predecir el número de etapas que aprobará un candidato, de modo que los candidatos mejor calificados sean detectados oportunamente para darles seguimiento de manera más inmediata.

1.4. Objetivos específicos

- Analizar y agrupar los datos sobre candidatos que el equipo de reclutamiento posee.
- Evaluar diferentes algoritmos y seleccionar el más adecuado para trabajar.

- Implementar y evaluar el algoritmo en los nuevos candidatos durante el proceso de reclutamiento.

1.5. Hipótesis

La implementación de un algoritmo de minería de datos con base en patrones de comportamiento ayudará a Nearsoft a identificar y priorizar a los candidatos que tendrán éxito a lo largo del proceso de reclutamiento.

1.6. Alcances y delimitaciones

La investigación se limitará a identificar el conocimiento explícito que posee el equipo de reclutamiento de Nearsoft sobre las calificaciones otorgadas a los procesos de reclutamiento de cada uno de los participantes.

El análisis tomará como base únicamente los registros existentes hasta 2015 como referencia para la evaluación de los diferentes algoritmos y la implementación del que se adecue mejor. La cultura de los candidatos se encuentra fuera de la predicción.

1.7. Justificación

El proyecto planteado se desarrolló dado a que la empresa Nearsoft reconoce que el proceso de reclutamiento es complejo y tienen un gran problema en la asignación de tiempo y recursos para cada uno de los candidatos en evaluación y han visto que esto ha causado la pérdida de candidatos valiosos que son contratados por otras empresas durante el proceso.

La implementación de un algoritmo beneficiará a la empresa dado a que, al identificar en una etapa temprana y agilizar su proceso, reducirá el riesgo de perder personal calificado a la hora de reclutar y ahorrará tiempo y esfuerzo al equipo de reclutamiento lo que se traduce en optimización de recursos, reducción de costos y contratación de personal de calidad.

2. MARCO DE REFERENCIA

En este capítulo se presenta la revisión literaria para profundizar en términos cuyas definiciones pueden resultar ambiguas. Inicialmente se expone una definición de los conceptos asociados a la presente investigación, tales como Recursos Humanos, Aprendizaje Automatizado y Minería de Datos; posteriormente se presentan los estúdios previos más comunes dentro de la minería de datos, seguidos de casos de éxito que, aunque no son iguales al caso presentado, sirvieron como guía para el desarrollo de la solución a las dificultades que se presentan en la empresa y finalmente los algoritmos más utilizados en la resolución de problemas como el que se presenta en este trabajo.

2.1. Recursos Humanos

El inicio de los recursos humanos se remonta a los problemas laborales de Estados Unidos, cuando las condiciones de trabajo eran extremadamente precarias dando lugar a huelgas, alta rotación de trabajo (Kaufman 2014). A principios del siglo XX apareció por primera vez un departamento de personal que tenía como objetivo mejorar las relaciones con los trabajadores manejando apropiadamente las quejas de los empleados, la seguridad y otros problemas del personal (DeNisi *et al.*, 2014).

El papel de los profesionales en recursos humanos en las organizaciones ha evolucionado paralelamente a los avances tecnológicos relacionados con las empresas. Los profesionales de recursos humanos están ahora en condiciones de dedicar más tiempo a las decisiones estratégicas de negocio dado a que el desarrollo de la tecnología ha permitido la automatización de procesos de recursos humanos. Aunque los profesionales de recursos humanos ya no son necesarios para el procesamiento manual de datos, no deben deslindarse de la tarea de recolección de datos sobre los empleados de la organización. Los datos sobre los recursos humanos que están disponibles dentro de una organización tienen el potencial de servir como ayuda en la toma de decisiones. El resto es identificar la información que es útil en las grandes bases de datos de recursos humanos que

son el resultado de la transacción de los procesos relacionados a recursos humanos (Wang, 2003).

2.2. Aprendizaje Automatizado

Tiempo atrás, el aprendizaje automatizado era conceptualizado como algoritmos que aprenden de forma autónoma a partir del contexto histórico específico de datos y basado en eso pueden hacer predicciones futuras con alta validez realizando tareas rutinarias y no rutinarias. El aprendizaje automatizado puede ser visto como una rama de la inteligencia artificial, ya que su núcleo se basa en software avanzado de reconocimiento de patrones para adaptarse a las nuevas circunstancias para detectar y extrapolar patrones (Mena, 2011; Strohmeier y Piazza, 2015; Kahraman, 2015).

En teoría, el aprendizaje automatizado de las máquinas podría detectar y extrapolar los patrones de los recursos humanos y los datos de la empresa y luego proporcionar una línea de administración en tiempo real y fiable sin participación de un profesional de los recursos humanos.

2.3. Minería de Datos

La minería de datos se refiere a extraer conocimiento de grandes cantidades de datos (Badr El Din Ahmed y Sayed Elaraby, 2014). Es la exploración de los datos históricos (por lo general de grandes cantidades) en busca de un patrón persistente y/o una relación sistemática entre las variables; se utiliza para validar los resultados mediante la aplicación de los patrones detectados a nuevos subconjuntos de datos (Giudici, 2003; Berry y Linoff, 1999). Las raíces de la minería de datos se originan en tres áreas: la estadística clásica, la inteligencia artificial (AI) y aprendizaje automatizado (Mena, 2011). Pregibon (Grenander y Miller, 1998) describe la minería de datos como una mezcla de las estadísticas, la inteligencia artificial, y la investigación de base de datos, y señaló que no era un campo de interés para muchos hasta hace poco.

Según Fayyad(Fayyad *et al.*, 1996) la minería de datos se puede dividir en dos tareas: tareas de predicción y tareas descriptivas. La descriptiva utiliza técnicas de asociación, agrupaciones, entre otras, para encontrar patrones ocultos en grandes conjuntos de datos y ayudar en la toma inteligente de decisiones. La predictiva construye modelos usando conjuntos de reglas, árboles de decisión, redes neuronales y vectores de soporte, por mencionar algunos, para predecir la clase de un nuevo conjunto de datos (Mishra *et al.*, 2014). El objetivo final es la predicción; por lo tanto, la minería de datos predictiva es el tipo más común de esta área y es la que cuenta con la mayoría de las aplicaciones de las empresas o de situaciones referentes a la vida.

La minería de datos predictiva tiene tres etapas, como se muestra en la figura 2.1.

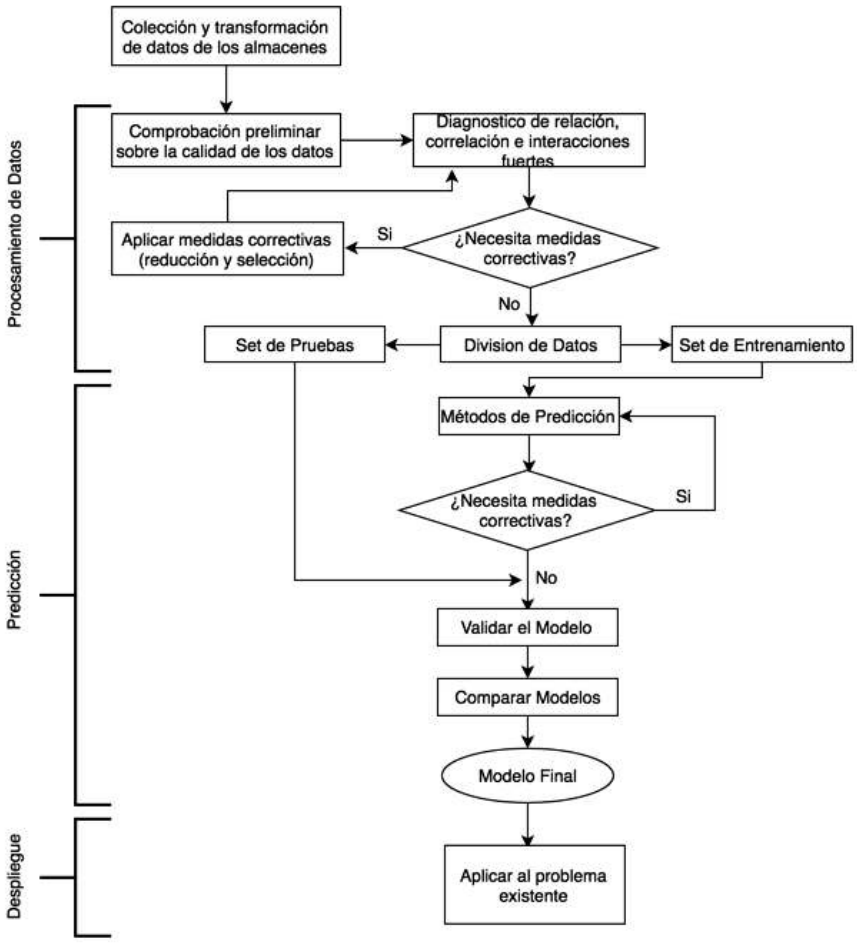

Figura 2.1 Etapas de la Minería de Datos

La minería de datos comienza con la recolección y almacenamiento de los datos en un almacén de datos. Estas dos principales tareas consisten en la identificación de las características relevantes de una empresa y el establecimiento de un archivo de almacenamiento para documentarlas. Es indispensable la limpieza y el aseguramiento de los datos para evitar su corrupción. Según Kimball, un almacén de datos es una copia de los datos transaccionales o no transaccionales estructurados específicamente para realizar consultas, análisis y presentación de informes (Kimball y Ross, 2002). La exploración de datos, que es el siguiente paso, puede incluir el análisis preliminar realizado a los datos para prepararlos para su análisis. El siguiente paso consiste en la selección de características y/o reducción, minar o construir modelos para la predicción es la tercera etapa, y finalmente llegan el post-procesamiento de datos, la interpretación y/o despliegue.

Las aplicaciones adecuadas para la minería de datos son enormes y todavía se están explorando en muchas áreas relacionadas a negocios y situaciones de la vida. Esto se debe a que, de acuerdo con Betts (2003), la minería de datos produce inesperadas pepitas de información que pueden abrir los ojos de una compañía a nuevos mercados, nuevas formas de llegar a los clientes y nuevas formas de hacer negocios. Por ejemplo, D. Bolka, Director del Departamento de Seguridad Nacional de Investigación Avanzada de Proyectos Agencia HSARPA (2004), según lo registrado por el IEEE de Seguridad y Privacidad (Goth, 2004), dijo que el concepto de minería de datos es una de esas cosas que se aplican a través del espectro, desde negocios que buscan en los datos financieros hasta los científicos en busca de nuevo conocimiento.

2.3.1. Clasificación y Predicción

Hoy en día existe una gran cantidad de datos que se recogen y almacenan en bases de datos de todo el mundo en todo el mundo. Hay información muy valiosa y conocimiento "oculto" en este tipo de bases de datos; y sin métodos automáticos para extraer esta información es prácticamente imposible minarlos.

La clasificación consiste en predecir un resultado determinado en base a una entrada dada. Con el fin de predecir el resultado, el algoritmo procesa un conjunto de entrenamiento que contiene un conjunto de atributos y el resultado respectivo, generalmente llamado meta o atributo de predicción. El algoritmo intenta descubrir las relaciones entre los atributos que harían posible predecir el resultado. A continuación, al algoritmo se le proporciona un conjunto de datos que no ha visto antes, llamado conjunto de predicción, que contiene el mismo conjunto de atributos, excepto por el atributo de predicción – aún desconocido. El algoritmo analiza la entrada y produce una predicción. La precisión de la predicción define la efectividad del algoritmo.

Para medir que tan bien hechas son las predicciones se compara el número de predicciones realizadas contra el número total de predicciones.

La minería de datos ofrece maneras prometedoras para descubrir patrones ocultos dentro de grandes cantidades de datos. Estos patrones ocultos potencialmente se pueden utilizar para predecir el comportamiento futuro de los datos.

2.3.2. Adquisición de Datos

En cualquier campo, el adquirir datos puede tomar una gran cantidad de esfuerzo y tiempo. Las lecturas y mediciones deben hacerse con instrumentos autónomos o ser capturados por operaciones comerciales en curso. Los instrumentos varían dependiendo del tipo de datos que se desea capturar, desde varios tipos de osciloscopios, multímetros, y libros de contabilidad. Hay una necesidad de registrar las mediciones y procesar los datos recogidos para su visualización, y esto se está volviendo cada vez más importante.

Se proporciona una descripción detallada del proceso de adquisición de datos de la investigación en cuestión en el capítulo 3.

2.3.3. Preparación de los Datos

Los datos son los ingredientes clave de todos los sistemas de aprendizaje supervisado. Pero los datos son inútiles por sí solos hasta que se extraen

conocimientos o inferencias de ellos. Casi todas las tareas de aprendizaje supervisado pueden ser formuladas como hacer inferencias sobre datos faltantes o latentes de los datos observados (Ghahramani, 2015).

Los datos brutos (por ejemplo, de un almacén) no siempre son los mejores para el análisis, y especialmente para la minería de datos predictiva. Los datos deben ser pre procesados o preparados y transformados para poderlos explotar de una manera óptima. La preparación de los datos es muy importante porque cada una de las técnicas de minería de datos predictivas se comportan de manera particular en función de la decodificación previa y métodos de transformación. Hay muchas técnicas para la preparación de datos que se pueden utilizar para lograr diferentes objetivos dentro de esta área.

2.4. Minería de Datos como herramienta en Recursos Humanos

Los recursos humanos son uno de los principales activos dentro de una organización para mejorar su ventaja competitiva en una economía del conocimiento. Entre las funciones de la gestión de recursos humanos, la selección de personal afecta de manera significativa el carácter de los empleados, por lo tanto es un tema importante para las organizaciones.

La minería de datos puede ayudar a los departamentos de recursos humanos a procesar de manera más eficiente los perfiles de candidatos a evaluar y poder extraer patrones de ellos para determinar cual tiene una mayor probabilidad de desempeñar de manera eficiente el trabajo para el cual aplica, especialmente en empresas grandes con mucha demanda de personal.

Este proceso puede considerarse como una aproximación a la evolución de análisis de datos en grandes bases de datos la cual podría convertirse en una herramienta útil para los profesionales de recursos humanos. La minería de datos consiste en la extracción de conocimiento a partir de los patrones de datos en bases de datos muy grandes. Sin embargo, la minería de datos va más allá de la realización de análisis de datos en grandes conjuntos de datos. Las

organizaciones que emplean a miles de empleados y el seguimiento de una multitud de información relacionada con el empleo podrían encontrar patrones valiosos de información contenida en sus bases de datos para proporcionar información en áreas tales como la retención de empleados y la planificación de compensaciones(Wang, 2003).

2.5. Trabajos previos

Existen varias aplicaciones conocidas de minería de datos como herramienta en el manejo y selección de recursos humanos, la mayoría de ellas utiliza o se basa en datos particulares de los candidatos para determinar el rendimiento de los postulantes, los resultados muestran que las técnicas aplicadas dan un resultado positivo en la mayoría de los casos.

Lin (2010) hizo una investigación en una compañía eléctrica y maquinaria en Taiwán, donde desarrolló una herramienta para la ayuda en la toma de decisiones mediante un proceso integrado a una red analítica.

En otra investigación (Singh Thakur *et al.*, 2015) se desarrolla un marco que permite a cualquier jefe de proyecto tomar la decisión correcta para la selección de nuevos talentos mediante la correlación de parámetros de rendimiento con otros atributos específicos del dominio de los candidatos. Además se comprueba la validez de otros procesos de selección los cuales se enfocan en el grado de estudio del candidato. Los resultados muestran que es necesario hacer un cambio a los criterios de selección que se usan normalmente para tener mejor calidad en el personal.

Los autores en (Gupta y V, 2013) han estudiado la importancia de las diferentes variables que juegan durante la selección de estudiantes, como promedio académico, habilidades de programación, dominio del conocimiento, habilidades de razonamiento, capacidad mental, habilidades matemáticas y más utilizando diferentes técnicas.

En (Chien y Chen, 2008) se hizo un trabajo sobre la mejora de la selección de personal en una industria de semiconductores, desarrollando un modelo, usando técnicas de minería de datos. Los atributos específicos incluían edad, género, estado civil, experiencia, educación, temas de interés y escolaridad como factores potenciales que pueden afectar el rendimiento.

Los autores de (Jantan *et al.*, 2010) investigaron varios factores que afectan el rendimiento de los empleados en el trabajo. Se revisaron trabajos previos que estudian el efecto de la experiencia, sueldo, capacitaciones, condiciones de trabajo y la satisfacción laboral en los parámetros de rendimiento.

La minería de datos soporta varias técnicas incluyendo análisis estadístico, árboles de decisión, algoritmos genéticos, clasificador de Bayes, técnicas de visualización, etc., para el análisis y la predicción.

2.6. Usos más comunes

Las técnicas de minería de datos conforman una rama de la inteligencia artificial, desde los años sesenta. Durante las décadas siguientes, las innovaciones importantes en los sistemas informáticos han conducido a la introducción de nuevas tecnologías(Ha *et al.*, 2000), para la educación basada en internet. La minería de datos permite una búsqueda para obtener información valiosa, en grandes volúmenes de datos(Weiss y Indurkhya, 1998). El crecimiento explosivo de las bases de datos ha creado una necesidad de desarrollar tecnologías que utilizan información y conocimiento de manera inteligente. Las técnicas de minería de datos se han hecho cada vez más importantes en la investigación(Fayyad *et al.*, 1996). A continuación se describen aplicaciones más comunes de las técnicas de minería de datos.

1. Redes Neuronales

El término red neuronal se utiliza tradicionalmente para referirse a una red o circuito de neuronas biológicas. El uso moderno del término se refiere a las redes neuronales artificiales, que se componen de neuronas artificiales, o nodos.

Además de la señalización eléctrica. Otras formas de señalización se derivan de difusión de transmisión neuronal, que tiene un efecto en la señalización eléctrica. Como tal, las redes neuronales son extremadamente complejas.

Algunas aplicaciones para redes neuronales son redes de función con base radial, clasificador neuronal, redes neuronales de propagación de confianza bayesianas, redes reguladoras de genes, redes neuronales recurrentes difusas, redes neuronales, redes neuronales artificiales de propagación, redes bayesianas, redes neuronales de regresión general y redes de flujo.

2. Arquitectura de Algoritmos

Las arquitecturas de algoritmo se expresan como una lista finita de instrucciones bien definidas, para el cálculo de una función. Los algoritmos se utilizan para el cálculo, procesamiento de datos y el razonamiento automatizado. En pocas palabras, un algoritmo es un procedimiento paso a paso para el cálculo. Una formalización parcial del concepto comenzó con los intentos de resolver el problema Entscheidungs, planteado por David Hilbert en 1928.

Algunas aplicaciones que se implementan por medio de algoritmos incluyen algoritmos de brecha estadística, automatización para la detección de la chi-cuadrada, modelos y algoritmos, arboles GRASP, OLAP, medias-K, algoritmos de agrupamiento, algoritmos de bosques de decisión, clasificación y regresión, la distancia euclídea, algoritmos de clúster, lógica difusa, regla de la asociación, algoritmo a priori, algoritmos genéticos, simulación de tiroides y SVM.

3. Enfoques Dinámicos basados en la Predicción

Es un modelo matemático para la dinámica estocástica; utilizado en el modelado de moléculas, en aplicaciones del mercado de valores y otras. La característica más importante de la dinámica de Langevin es la presencia de un ruido aleatorio gaussiano. El principio de localidad temporal fue ampliamente estudiado y aplicado en diferentes ámbitos de ciencias de la computación por Peter en 1968.

Se observa que un hilo no tiene acceso a un conjunto completamente aleatorio a través de sus transacciones.

Algunas aplicaciones que utilizan este enfoque incluyen la oncología oftálmica, diagnóstico de avería de un vehículo, computación, búsqueda previa, modelos de predicción de restauración de fallos, modelos de predicción de fallos, modelos de predicción de desastre financiero, ecuaciones de Maxwell-Vlasov, predicciones de reactividad química, seguimiento en tiempo real de un vehículo, detección de anomalías, rotación de predicción y predicciones clínicas.

4. Análisis de Sistemas de Arquitectura

El análisis de sistemas de arquitectura utiliza un modelo conceptual que define la estructura, el comportamiento y otros aspectos de un sistema. Los sistemas de arquitectura hacen uso de elementos de software y hardware, lo que permite el diseño de sistemas compuestos. Una buena arquitectura puede ser vista como un “esquema de partición” o algoritmo que divide completamente todos los requisitos actuales y previsibles del sistema en un conjunto viable de subsistemas claramente delimitadas.

Algunas aplicaciones de los análisis de sistemas de arquitectura son análisis semántico, análisis de regresión, análisis estadístico, análisis discriminante, análisis de asociación, análisis discriminante penalizado, análisis de parámetros de proceso, análisis de conglomerados, toma de decisiones, sistemas de ayuda en toma de decisiones, análisis de comportamiento de consumidores, análisis de regresión logística binaria, árboles de modelo M5, análisis factorial, análisis de mercado, filtrado colaborativo, análisis de datos, modelos basados en árboles de decisión, análisis de componentes principales, selección multi-función, detección de intrusos y hemodiálisis.

5. Sistemas de Agentes Inteligentes

En el campo de la Inteligencia Artificial, un sistema de agentes inteligentes es una entidad autónoma, que observa y actúa sobre un medio ambiente. Los agentes

inteligentes también pueden aprender, o utilizar el conocimiento para alcanzar sus objetivos que pueden ser muy simples o muy complejos. Una máquina de reflejo, tal como un termostato es un agente inteligente, como un ser humano, en una comunidad de seres humanos que trabajan juntos hacia un objetivo.

Algunas aplicaciones para sistemas de agentes inteligentes incluyen sistemas multi lenguaje, sistemas complejos, diseño de interfaz de la computadora, sistemas de bases de datos multiusuarios, análisis inteligente, inteligencia de fabricación, sistemas inteligente de tutores, máquinas de vectores, sistemas de diagnóstico, sistemas de control y especializados e inteligencia artificial.

6. Modelado

El modelado, en la ingeniería de software, es el proceso de crear un modelo de datos al hacer descripciones de modelos de datos formales, utilizando técnicas de modelado de datos. La tecnología de modelado puede proporcionar métodos cuantitativos para el análisis de datos, para representar o adquirir un conocimiento experto, utilizando la programación lógica inductiva, o algoritmos, por lo que la inseminación artificial, la ciencia cognitiva y otros campos de investigación producen plataformas más amplias para el desarrollo de técnicas de minería de datos.

Las aplicaciones de modelado incluyen el modelado de costos, diagnóstico basado en modelo, modelado de la proliferación de incendios forestales, estadísticas de salida, modelado de la entonación, modelado de documento XML, modelado de riesgos proporcionales de Cox, exponentes de daños de carga, polinomios, formas de onda similares, peso de un aditivo simple, control numérico por computadora, aprendizaje y utilización de drogas.

7. Sistemas basados en conocimiento

Los sistemas basados en conocimiento son herramientas de Inteligencia Artificial que trabajan en un dominio estrecho, para proporcionar decisiones inteligentes, con justificación. La definición más común de sistemas basados en conocimiento

está centrada en el hombre, ya que tiene sus raíces en el campo de la inteligencia artificial.

El conocimiento es adquirido y representado, utilizando varias técnicas de representación de conocimiento, reglas, normas y scripts. Las ventajas básicas ofrecidas por estos sistemas son la documentación de los conocimientos, ayuda a la decisión inteligente, auto-aprendizaje, el razonamiento y la explicación.

Algunas aplicaciones de los sistemas basados en el conocimiento incluyen técnicas de aprendizaje, técnicas de control automático, descubrimiento de conocimiento en bases de datos, espirales de conocimiento, tecnologías de comunicación, mediciones de conocimiento, extracción de conocimiento, adquisición de conocimiento, gestión del conocimiento, representación del conocimiento, bibliotecas digitales e información de ganancia de minería de datos.

8. Sistemas de Optimización

Fermat y LaGrange utilizaron fórmulas basadas en cálculos, para la identificación óptima, mientras que Newton y Gauss propusieron métodos iterativos de acercarse a un óptimo. Históricamente, el término original para la optimización era “programación lineal”, acuñado por George B. Dantzig, aunque gran parte de la teoría había sido descrita por Leonid Kantorovich, en 1939.

Dantzig publicó el algoritmo Simplex, en 1947 y desarrolló la teoría de la dualidad, en el mismo año. La optimización del sistema se refiere a la selección de un mejor elemento, a partir de un conjunto de alternativas disponibles. En el caso más simple, problemas en los que se maximiza una función real, se resuelven mediante elección sistemática de los valores de las variables reales o enteros, desde dentro de un conjunto permitido.

Algunas aplicaciones incluyen estimulación eléctrica de los nervios, detección del pico R, métodos de extracción de referencia individuales latentes, los valores de optimización de operación, partición vertical, regresión logística, procesos de jerarquía analítica, regresión polinómica, optimización basada en la biogeografía,

optimización de enjambre de partículas, los métodos de elementos finitos, métodos discretos de aproximación, métodos asintóticos y la computación paralela.

9. Sistemas de información

Los sistemas de información son el producto de una disciplina académica. Ocupan un lugar entre el mundo de los negocios y la informática, haciendo el puente entre el campo de los negocios y el bien definido campo de la informática que está evolucionando hacia una nueva área científica. Un sistema de información se basa en los fundamentos teóricos de la información y la informática, que permite a los investigadores una oportunidad única para participar en los estudios académicos de diversos modelos de negocio y procesos algorítmicos relacionados que tengan relación la informática.

Algunas aplicaciones de los sistemas de información incluyen las características del paciente, bases de datos móviles, auto organización, mapas de características, bases de datos de reclamaciones de seguros, alteración de un campo de corriente, acidificación, sub-series de tiempo, elección de destino, atributo de los estudios de relevancia, estados financieros fraudulentos, la similitud de secuencias, razonamiento basado en casos, los datos antropométricos, Splines de regresión, los desequilibrios económicos, los fallos de los clientes de media tensión, mantenimiento e ingeniería, los préstamos bancarios, el aprendizaje por refuerzo, supervisión de aprendizaje, visualización de información, retención de clientes, gestión de rotación, patrón de descubrimiento, gestión de relaciones con los clientes y la resistencia a la compresión uniaxial.

2.7. Casos de éxito

Desde su creación, el campo de la minería de datos y descubrimiento de conocimiento a partir de bases de datos ha sido impulsado por la necesidad de resolver problemas prácticos. Desde manejar grandes bases de datos con datos imprecisos, hasta buscar patrones de asociación de datos en las transacciones de los supermercados, la minería de datos es un área de investigación rica en

aplicación. A pesar de sus raíces prácticas se han publicado algunos estudios de caso de las aplicaciones de minería de datos (Melli *et al.*, 2006).

Una aplicación de minería de datos para la mejora de la calidad de la atención y la reducción de los gastos sanitarios que rodea la enfermedad cardiovascular fue desarrollada por los doctores del Grupo de Soluciones Medicas Siemens(Rao *et al.*, 2006). En una nota práctica, el artículo presenta una nueva utilización de los conocimiento de dominio externo para solucionar el problema frecuente de los datos incompletos y ruidosos, resultando en una mejora de diagnóstico salvando la vida de varios pacientes.

La aplicación de la minería de datos se ha aplicado también con éxito en el área gubernamental y financiera. En Estados Unidos, la implementación de un sistema de minería de datos para uso en el Servicio Interno de Impuestos logró identificar a las personas con altos ingresos que participan en la evasión de impuestos, reconociendo cientos de millones de dólares en ingresos protegidos, incluyendo la emisión de amparos (DeBarr y Eyler-Walker, 2006).

En(Agarwal et al., 2006) se aborda el problema prevalente de la identificación de sitios web con contenido inaceptable como la pornografía. El artículo presenta una representación ampliada de la información codificada y relacionada con una página web con el fin de mejorar la clasificación y demostrar su rendimiento en AOL. A juzgar por los resultados experimentales, las soluciones propuestas son prometedoras.

Dentro del área de la mercadotecnia podemos ver que Freeman y Melli (2006) revisan algunos de los obstáculos organizativos que la aplicación de un sistema de minería de datos puede encontrar dentro de una gran corporación. La aplicación específica es un modelo predictivo para el valor de tiempo de vida de cliente en una compañía de telecomunicaciones. El artículo será de especial interés para los expertos en minería de datos que también son requeridos para defender con éxito un proyecto de minería de datos.

El proceso de desarrollo de software también puede ser un caso de estudio, dado a que en (Kanellopoulos *et al.*, 2006) se puede encontrar la aplicación de minería de datos para el proceso de mantenimiento de software orientado a objetos. El desarrollo de sistemas de software complejo fiable es clave para la ingeniería de software. El enfoque utilizado implica el agrupamiento de código fuente con el fin de recuperar el conocimiento acerca de la estructura de mantenimiento de un software (mucho tiempo de mantenimiento es empleado en la comprensión del programa). El artículo incluye un enfoque novedoso e integral a la validación empírica.

Otro ejemplo, esta vez orientado a la industria del por menor, se puede encontrar la investigación de la aplicación de minería de datos para la etapa de preparación de datos en modelos de predicción (Ghani *et al.*, 2006). Uno de los desafíos que enfrenta el practicante es este dominio, si la tarea es previsión de la demanda, optimización de surtido o recomendación de productos, es la extracción de atributos relevantes. En particular, el interés actual es la extracción de atributos implícitos.

Se realizó un estudio en una compañía de seguros con minería de datos, utilizando diversos métodos para el análisis de las causas fundamentales de las quejas de los clientes hacia la compañía (Ha y Park, 2006). Describe una implementación para una compañía de seguros de vida que localiza dónde sucedieron las quejas, la relación entre los problemas y la causa de los problemas. El problema es importante y no trivial ya que las reacciones de los clientes son independientes de la arquitectura interna de la empresa.

En (Singh *et al.*, 2006) se presenta la aplicación del análisis de la cesta en las ventas cruzadas de la tienda en línea y el centro de llamadas de la tienda Hewlett-Packard. El documento ilustra los retos que enfrenta comúnmente al realizar una prueba de concepto en un marco de tiempo muy corto, como requisito previo para la financiación de la implementación de sistemas.

Por último, se puede encontrar una aplicación de minería de datos para la predicción de clientes que pueden estar asociados con eventos de morosidad, tales como la falta de pago (Pnheiro *et al.*, 2006). Este tema ha recibido interés por parte de ambas empresas de servicios y proveedores de minería de datos, pero son pocos los casos de estudio que se han publicado hasta la fecha. El documento abarca varias de las fases en el proceso de minería de datos que incluyen una fase de segmentación de clientes.

2.8. ¿Por qué no es suficiente o qué vamos a mejorar?

De acuerdo con la literatura, los modelos más utilizados para aprendizaje supervisado son los árboles de decisión, sin embargo en esta investigación no se pretenden utilizar datos sociodemográficos (género, edad, etc.) u otros datos de desempeño como la escolaridad o el promedio. La propuesta presentada quiere explorar nuevos enfoques para encontrar respuestas incorrectas que son elegidas con más frecuencia, incluso más que las respuestas correctas y que se ignora el porqué.

2.9. Comparativa de los algoritmos para Minería de Datos

Dentro de la minería de datos, existe una diversa cantidad de algoritmos de aprendizaje supervisado para la clasificación. En esta investigación se explorarán principalmente tres, los cuales son Árboles de Decisión, Ingenuo de Bayes y Máquinas de Vectores de Soporte dado a que son los más comunes y su implementación no es complicada.

Cada uno de los algoritmos antes mencionados, tiene una metodología diferente para determinar si una instancia pertenece a una clase, y dependiendo del problema a resolver, el desempeño puede variar. En este trabajo se comparan los resultados de la aplicación de estos tres métodos o algoritmos para la selección de personal asistido por un método de Machine Learning.

2.9.1. Naïve Bayes

Dado un conjunto de objetos, cada uno de los cuales pertenece a una clase conocida, y cada uno de los cuales tiene un vector conocido de variables, el objetivo es construir una regla que permita asignar objetos futuros a una clase, teniendo en cuenta sólo los vectores de variables que describen los objetos futuros. Los problemas de este tipo, llamados problemas de clasificación supervisada, están en todas partes, y se han desarrollado muchos para construir tales normas. Uno muy importante es el método de Ingenuo de Bayes. Este método es importante por varias razones, es muy fácil de construir, no necesita ningún esquema de estimación con parámetros iterativos. Esto significa que puede aplicarse fácilmente a grandes conjuntos de datos. Es fácil de interpretar, así los usuarios sin mucha habilidad en tecnologías de clasificadores pueden entender porque el clasificador hace lo que hace. Y finalmente, a menudo funciona sorprendentemente bien; puede que no sea el mejor clasificador posible para una aplicación en particular, usualmente puedes confiar en que va a ser robusto y funcionar correctamente.

El principio básico

Si se asumen solo dos clases, etiquetadas como $i = 0{,}1$. El objetivo es tener dos clases, donde la más grande tenga la etiqueta 1 (conjunto de entrenamiento) y la clase más pequeña es la que tenga los objetos etiquetados con 0 (conjunto de pruebas). La clasificación se consigue entonces mediante la comparación de este marcador con un umbral, t. Si definimos $P(i|x)$ como la probabilidad de que un objeto con vector de medición $x = (x_1, \ldots x_p)$ pertenece a la clase i, entonces cualquier función monótona de $P(i|x)$ tendría una puntuación adecuada. En particular, la relación $P(1|x)/P(0|x)$ sería adecuada. La probabilidad elemental nos dice que podemos descomponer $P(i|x)$ como proporcional a $f(x|i)P(i)$, donde $f(x|i)$ es la distribución condicional de x para i objetos, y $P(i)$ es la probabilidad de que un objeto pertenece a la clase i si no sabemos nada más sobre él (la probabilidad “a priori” de la clase i). Esto significa que la relación se vuelve:

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)}$$

Para utilizar esto para producir clasificaciones, se necesita estimar la $f(x|i)$ y la $P(i)$. Si el conjunto de entrenamiento es una muestra aleatoria de la población general, $P(i)$ se puede estimar directamente de la proporción de objetos de la clase i en el conjunto de entrenamiento. Para estimar la $f(x|i)$, el método Ingenuo de Bayes asuma que los componentes de x son independientes, $f(x|i) = \prod_{j=1}^{p} f(x_j|i)$, y luego se estiman cada una de las distribuciones univariantes $f(xj|i)$, $j = 1,\ldots p;\ i = 0,1$, por separado. Así, el problema de dimensiones multivariables p se ha reducido a problema de estimación univariante p. La estimación univariante es familiar, simple y requiere un tamaño menor del conjunto de entrenamiento para obtener estimaciones exactas. Esta es una de las ventaja del método de Ingenuo de Bayes: la estimación es simple, muy rápida y no requiere esquemas complicados de interacciones.

2.9.2. Máquinas de Vectores de Soporte

Las máquinas de Vectores de Soporte (SVM) un poderoso, robusto y sofisticado método de aprendizaje supervisado. Se basa en la teoría del aprendizaje estadístico. Fue propuesto por primera vez por Cortés y Vapnik de su trabajo original sobre minimización del riesgo estructural y luego modificado por Vapnik (Kanchanamani y Perumal, 2016).

En las aplicaciones de aprendizaje automatizado de hoy en día, las SVM (Vapnik, 1995) son consideradas básicas para probar, dado a que ofrecen uno de los métodos más sólidos y precisos entre los algoritmos mejor conocidos. Tiene una base teórica sólida, requiere solamente una docena de ejemplos para la formación, y es insensible al número de dimensiones. Además, métodos eficientes para entrenar las SVM están siendo desarrollados a gran velocidad.

En una tarea de aprendizaje de dos clases, el objetivo de SVM es encontrar la mejor función de clasificación para distinguir entre los miembros de las dos clases

en los datos de entrenamiento. La métrica para el concepto de la función de más óptima clasificación se puede realizar geométricamente. Para un conjunto de datos linealmente separables, una función de clasificación lineal corresponde a una separación hiperplana $f(x)$ que pasa por el medio de las dos clases, separando a ambas. Una vez determinada esta función, una nueva instancia de datos x_n puede ser clasificada simplemente probando el signo de la función $f(x_n)$; donde xn pertenece a la clase positiva si $f(x_n) > 0$.

Debido a que hay muchos hiperplanos lineales, SVM garantiza que la mejor función es encontrada maximizando el margen entre las dos clases. Intuitivamente, el margen es definido como la cantidad de espacio o separación entre las dos clases definidas en el hiperplano. Geométricamente, el margen corresponde a la distancia más corta entre los puntos de datos más cercanos a un punto en el hiperplano. Teniendo esta definición geométrica, nos permite explorar cómo maximizar el margen, por lo que a pesar de que hay un número finito de hiperplanos, solo unos pocos califican como una solución a SVM.

La razón por la cual SVM insiste en encontrar el margen máximo de los hiperplanos es que ofrece la mejor capacidad de generalización. Permite no solo la mejor clasificación de rendimiento (por ejemplo, precisión) en los datos de entrenamiento, sino que también deja mucho espacio para la correcta clasificación de los datos futuros. Para asegurarse de que realmente encuentre el margen máximo de los hiperplanos, el clasificador SVM intenta maximizar la siguiente función con respecto a $\vec{w}$ y b:

$$L_P = \frac{1}{2}\left|\left|\vec{w}\right|\right| - \sum_{i=1}^{t} a_i y_i (\vec{w}\,\vec{x}_i + b) + \sum_{i=1}^{t} a_i$$

donde t es el número de ejemplos de entrenamiento, y $a_i, i = 1, \ldots, t$, son números no negativos como los derivados de L_P con respecto a a_i son cero. a_i son los multiplicadores de Lagrange y L_P es llamado el Lagrangiano. En esta ecuación, los vectores $\vec{w}$y la constante b definen el hiperplano.

3. MODELO

En este capítulo se realiza la propuesta de modelo que se busca llevar a cabo con la finalidad de lograr para predecir el éxito de los candidatos. Como resultado de la primera etapa de esta investigación, consistente en la evaluación de técnicas y estrategias para la solución de este problema, se identificaron metodologías para la minería de datos como Naïve Bayes, SVM y Árboles de decisión.

En relación al tipo de investigación del presente estudio, este posee un enfoque mixto. (Hernandez Sampieri *et al.*, 2014) indica que los métodos mixtos representan un conjunto de procesos sistemáticos, empíricos y críticos de investigación e implican la recolección y el análisis de datos cuantitativos y cualitativos, así como su integración y discusión conjunta, para realizar inferencias producto de toda la información recabada (meta inferencias) y lograr un mayor entendimiento del fenómeno bajo estudio.

Esta no es una investigación de corte cuantitativo ni cualitativo ya que se basa en el análisis de datos y el diseño de aplicaciones basadas en minería de datos utilizando algoritmos de Inteligencia Artificial para la detección de patrones que son imperceptibles para el humano. El modelo consiste en analizar datos, pre procesarlos y alimentar modelos de Inteligencia Artificial de métodos supervisados de aprendizaje que detectarán y aprenderán los patrones y características de las clases (en este caso, 2 clases: pasar la etapa de evaluación o no) y serán capaces de clasificar una muestra nunca antes vista con un alto porcentaje de acierto.

En la revisión se utilizan diferentes procesos en estudios previos los cuales no son descritos a profundidad; sin embargo, en la mayoría de ellos se siguen una serie de pasos generales los cuales se encuentran descritos en el modelo de la figura 3.1.

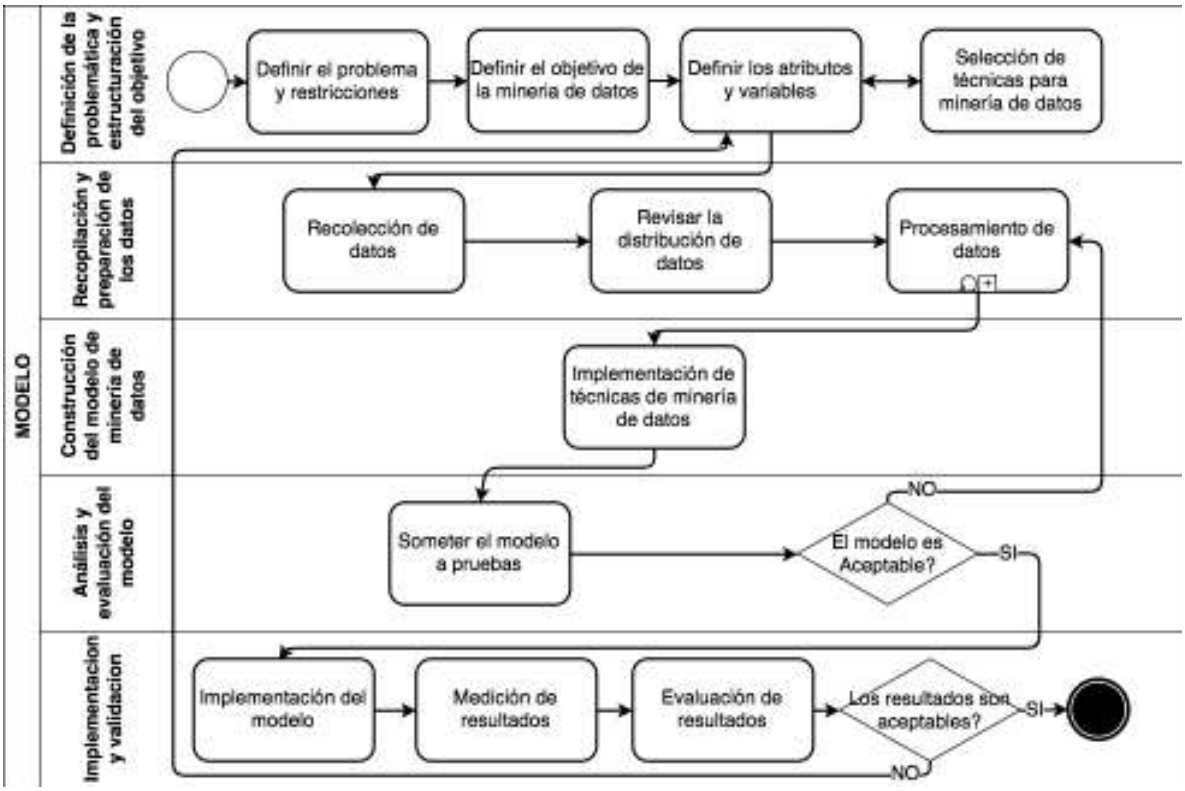

Figura 3.1 Modelo para minería de datos.

La figura 3.1 muestra el modelo generado con base en las implementaciones examinadas en el capítulo anterior, el cual se compone de 5 etapas:

A. Definición de la problemática y estructuración del objetivo.
B. Recolección y preparación de los datos.
C. Construcción del modelo de minería de datos.
D. Análisis de evaluación del modelo.
E. Implementación y validación.

Cada una de estas etapas, consta de tareas independientes que serán descritas a continuación.

3.1. Definición de la problemática y estructuración del objetivo.

El primer paso es entender y definir el problema correctamente y especificar los objetivos para la minería de datos. Mientras tanto, las personas encargadas para minar los datos deben tener dominio tanto de técnicas como de la problemática

para poder entender la naturaleza del problema, lo que mejorará en gran medida la efectividad y eficiencia de la minería de datos.

3.1.1. Definir el problema y las restricciones

Para definir el problema es necesario entender perfectamente la problemática de la empresa y el impacto que conlleva el hecho de que no sea resuelta. Es importante también tener presente cada una de las restricciones necesarias para abordar el tema, ya que de alguna manera afectan la realización del estudio. Aquí se debe anotar todas las limitaciones que se pueden presentar en el desarrollo de la investigación, teniendo en cuenta hasta el final de la investigación.

3.1.2. Definir el objetivo de la minería de datos

La minería de datos es una forma innovadora de obtener información valiosa mediante el análisis de los datos contenidos en una base de datos. Esta información sirve de ayuda para una adecuada toma de decisiones empresariales. Esencialmente, la minería de datos es un método innovador de aprovechar la información ya existente a fin de, por ejemplo, mejorar procesos, predicción de eventos, detectar patrones y encontrar conocimiento en los datos.

3.1.3. Definir los atributos y variables

Un atributo es una característica de un objeto y están estrechamente relacionados con las variables. Una variable es un conjunto lógico de atributos. Las variables pueden "variar" - por ejemplo, ser alta o baja. Qué tan alto, o tan bajo, está determinado por el valor del atributo (y de hecho, un atributo podría ser simplemente la palabra "bajo" o "alto").

Mientras que un atributo es a menudo intuitivo, la variable es la manera operacionalizada en la cual el atributo se representa para el procesamiento adicional de datos. En el procesamiento de datos, los datos suelen estar representados por una combinación de elementos (objetos organizados en filas) y múltiples variables (organizadas en columnas).

Para la problemática planteada, las variables se pueden definir en función de datos tanto de los candidatos como de la entrevista. Por ejemplo los datos de los candidatos pueden ser edad, género, escolaridad, estado civil, etc., mientras que los datos de la entrevista pueden ser cada una de las preguntas realizadas o el resultado de las evaluaciones dada la experiencia del postulante.

3.1.4. Seleccionar técnicas de minería de datos

Esta tarea consiste en analizar las diferentes técnicas para minería de datos que existen en la actualidad, evaluar la complejidad de cada una de ellas y seleccionar la que mejor se adapte a la problemática planteada para una resolución más eficiente.

Dentro de los algoritmos que se pueden utilizar se encuentra el de Naïve Bayes, el cuál es el más utilizado en la actualidad debido a su sencillez y aun así es de los más potentes; este algoritmo trabaja de manera eficaz cuando las variables de la investigación son independientes. Por otro lado el algoritmo de SVM es muy eficiente ya que maneja un enfoque diferente porque ubica muestras en un espacio multi dimensional. Es importante comparar los resultados de los algoritmos seleccionados contra los más utilizados en el área, por ejemplo los árboles de decisión.

3.2. Recolección y preparación de los datos.

A través de la comprensión de las fuentes y tipos de datos relacionados que se pueden recopilar, hacerlo de manera correcta es la base de la minería de datos. De hecho, los datos de los recursos humanos deberían almacenarse en una base de datos independiente por privacidad.

Los datos relacionados deben combinarse y prepararse antes de un análisis posterior. Sin embargo, los datos recogidos a menudo incluyen datos ruidosos, ausentes e incoherentes.

3.2.1. Recopilación de los datos

La recopilación de datos es el proceso de recolección y medición de información sobre las variables seleccionadas de manera sistemática y establecida, lo que permite responder a las preguntas pertinentes y evaluar los resultados. El componente de recolección de datos de investigación es común a todos los campos de estudio, incluyendo las ciencias físicas y sociales, las humanidades y los negocios. El objetivo de toda la recolección de datos es capturar evidencia de calidad que luego se traduce en análisis de datos ricos y permite construir una respuesta convincente y creíble a las preguntas que se han planteado.

Generalmente hay cuatro tipos de recolección de datos y son:

1. Encuestas: Estándar de papel y lápiz o cuestionarios telefónicos que hacen preguntas predeterminadas.
2. Entrevistas: conversaciones individuales estructuradas o no estructuradas con individuos o líderes clave en una comunidad.
3. Grupos focales: Entrevistas estructuradas con pequeños grupos de individuos similares usando preguntas estandarizadas, preguntas de seguimiento y exploración de otros temas que surgen para entender mejor a los participantes.
4. Bases de datos: conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

3.2.2. Revisar la distribución de los datos

Una recopilación de datos se puede distribuir o dispersar de muchas maneras diferentes. Por ejemplo, los datos de lanzar un dado pueden variar uniformemente de 1 hasta 6. Los datos de un proceso de manufactura pueden estar centrados en un valor objetivo. Los datos de la industria de la salud pueden incluir valores que se encuentren muy alejados del valor central.

Se puede evaluar una distribución de datos por medio de gráficas, estadísticos descriptivos o una herramienta de identificación de distribución:

A. Evaluaciones visuales

Las gráficas como los histogramas pueden proveer instantáneamente información sobre la distribución de un conjunto de datos. Los histogramas pueden ayudarle a observar:

1. Si los datos se agrupan en torno a un valor individual o si los datos tienen múltiples picos o modas.
2. Si los datos están diseminados con poca densidad en un rango amplio o si los datos se encuentran dentro de un rango pequeño.
3. Si los datos son asimétricos o simétricos.

B. Métricas de las distribuciones

Los estadísticos descriptivos que describen la tendencia central (media, mediana) y la dispersión (varianza, desviación estándar) de los datos con valores numéricos agregan un nivel de detalle y se pueden utilizar para hacer comparaciones con otros conjuntos de datos.

C. Definiciones formales o teóricas

Finalmente, algunas distribuciones comunes se pueden identificar como distribución normal, exponencial y de Weibull (Gorgoso *et al.*, 2007). Conocer la distribución de un conjunto de datos puede proporcionar información sobre los datos en sí y puede ser crítico al seleccionar los análisis apropiados e interpretar los resultados.

3.2.3. Procesamiento de datos

Este proceso involucra una serie de subprocesos para el procesamiento de los datos. Es un proceso cíclico que conlleva actividades importantes de los datos tales como la transformación, revisión, reducción y enriquecimiento. Es importante que sea un proceso cíclico dado a que es aquí el punto de partida para la implementación de las técnicas de minería de datos.

Los pasos están representados en la figura 3.2 y se describen a continuación.

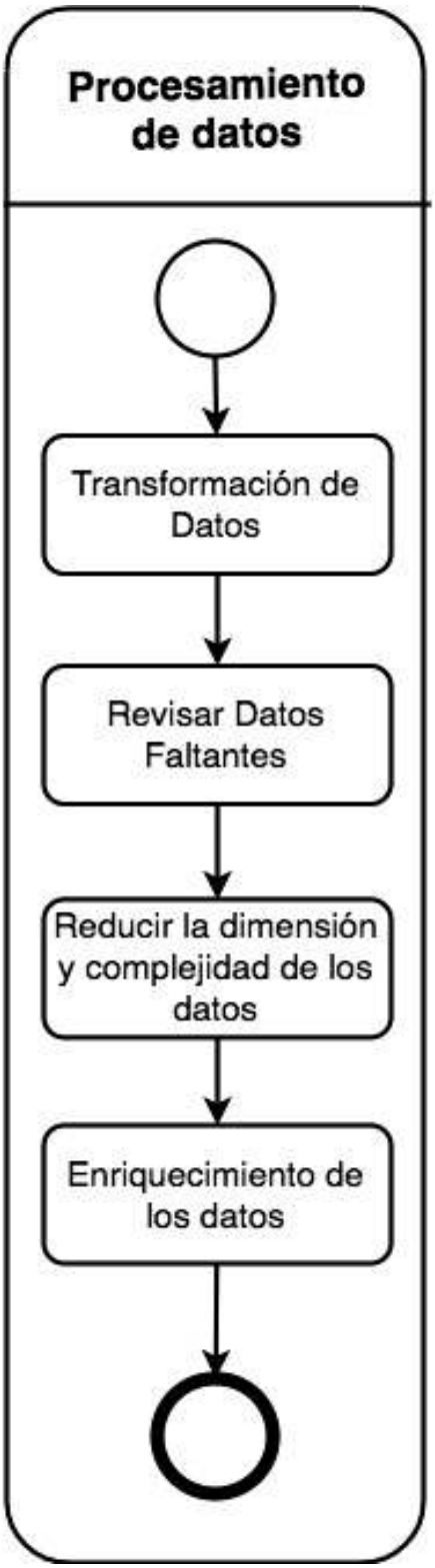

Figura 3.2 Ciclo de procesamiento de datos.

A. Transformación de los datos

En este apartado, se deben revisar los datos para determinar si se encuentran en un tipo manejable, por ejemplo valores binarios, números enteros, texto, etc. De no ser el caso, se debe identificar cuáles de los valores necesitan ser transformados a un tipo de mayor conveniencia. De igual manera, si los datos se encuentran en una base de datos, distribuidos en tablas, es importante encontrar la relación entre ellos y si es posible, normalizar la estructura para un mejor manejo.

B. Revisar datos faltantes

Es esencial identificar las causas de la falta de datos para evitar que esos problemas con los datos no vuelvan a ocurrir. Las soluciones para datos faltantes incluyen rellenar manualmente los valores perdidos y rellenar automáticamente con la palabra “vacio”, “desconocido”, “nulo”, etc.

Se debe prestar suma atención a este proceso, ya que en caso de que una serie de datos importantes no se encuentren presentes, es casi imposible poder tener un resultado fiable después de la implementación de minería de datos.

Es importante revisar que cada uno de los datos necesarios para la resolución de la problemática se encuentre presente.

C. Reducir la dimensión y complejidad de los datos

En la mayoría de las ocasiones, existe información dentro de los datos que no es importante para la minería o no se va a tomar en cuenta para el minado dado a que así fue establecido en un inicio. Es indispensable reducir la cantidad de datos para limitar la base a únicamente datos útiles, lo que va a elevar la calidad de los mismos.

Existen diferentes tipos de técnicas para lograr esta tarea, la cual también es conocica como normalización de base de datos. Es un proceso donde se descartan los datos repetidos y se eliminan redundancias.

D. Enriquecimiento de los datos

Se refiere a la implementación de un conjunto de métodos y estructuras para refinar y mejorar la información. Más concretamente, el objetivo de enriquecer los datos es convertirlo en un activo más valioso: sacarle más provecho, hacer más con él, acceder a él más fácilmente y ser más proactivo en su uso, sin aumentar notablemente los costos ni riesgos.

3.3. Construcción del modelo de minería de datos

El problema actual de la predicción de éxito en los candidatos puede estructurarse como un problema de clasificación. Teniendo en cuenta las necesidades de interpretación de resultados y justificación de reglas, se emplea un árbol de decisiones para la minería de datos.

Una práctica común es separar los datos desde un inicio, tomando el mayor porcentaje de ellos para crear el modelo de predicción y utilizando el resto para comprobarlos. De esta manera, se pueden hacer pruebas sobre el modelo resultante con datos los cuales ya se conoce el comportamiento de ante mano, sin necesidad de implementar directamente en la empresa. Lo recomendable es separar 90% de los datos para la construcción del modelo y reservar el 10% restante para pruebas.

3.3.1. Implementación de técnicas de minería de datos

A. Naïve Bayes

Una vez que se tienen los datos, es importante clasificar los posibles eventos e identificar las variables que están involucradas para que estos resultados sean dados. Ya que se han identificado estos eventos y variables, se debe entrenar un modelo probabilístico el cual va a servir para predecir un nuevo evento dado un set de variables introducidas.

Se establece el procedimiento establecido por el clasificador Naïve Bayes, el cual conlleva los siguientes pasos:

1. Construcción de un conjunto de datos para entrenamiento, el cual consiste en las respuestas de los candidatos al examen técnico. Estos registros contienen los datos especificados en la figura 4.1.
2. Se prepararán cada uno de los casos tanto positivos como negativos relacionados al resultado de todas las etapas de entrevista, separándolos en distintas clases.
3. Para cada una de las clases se calcula la probabilidad de que un valor o respuesta exista en cada uno de los reactivos.
4. Para la prueba de efectividad del modelo se toma el 10% de los registros existentes, y para cada uno se calcula la probabilidad de pertenecer a cada clase, esto es aprobar cada una de las etapas o ser descalificado.
5. El porcentaje de aciertos de la predicción del modelo es el criterio para medir su efectividad (apartado 3.4).

B. SVM

Los datos se consideran como un conjunto de puntos en los que cada uno pertenece a dos posibles categorías, el modelo construido basado en este algoritmo es capaz de predecir si un nuevo dato pertenece a una categoría u otra. Los datos de entrada son vistos como puntos en un vector p-dimensional. SVM busca un hiperplano que separe de forma óptima a los puntos de ambas clases.

Estas técnicas de minería de datos pueden ser implementadas de diferentes maneras, las mas comunes son a través de software especializado.

3.4. Análisis de evaluación del modelo.

El modelo construido debe ser revisado y evaluado antes de que pueda utilizarse. Para evaluar el mismo, se deben someter los datos de prueba para verificar la exactitud con la que puede predecir los resultados.

3.4.1. Someter el modelo a pruebas

Una vez que se tiene un modelo como resultado de la implementación de minería de datos, se debe someter a pruebas para validación antes llevarlo al entorno real de trabajo.

Dado a que el modelo resultante es entrenado a partir del 90% de datos existentes, se puede utilizar el 10% restante para probar la exactitud del mismo. De estos últimos, se conoce previamente el comportamiento y resultado, por lo que el algoritmo emitirá una predicción sobre los datos ya conocidos (aprobar o no una determinada etapa del proceso). El criterio de evaluación del desempeño de nuestra propuesta será calculado sobre el porcentaje de aciertos del conjunto de datos de pruebas.

Para comprobar la consistencia en la predicción de la solución se puede utilizar la técnica de validación cruzada, donde los conjuntos de prueba y entrenamiento se eligen de manera totalmente aleatoria en n ocasiones y el desempeño en la predicción es un promedio de los resultados de cada iteración.

Para el caso particular de este modelo, dado a que el resultado de cada etapa son dos posibles valores (pasar y no pasar, con una probabilidad de 50% cada uno), se puede considerar que el mismo tiene éxito si el porcentaje de aciertos en las pruebas es mayor a la probabilidad aleatoria de obtener cual sea de estos dos valores.

En el caso de que el modelo no sea efectivo se deben realizar varias actividades. La primera de ella es evaluar los datos en los que el modelo no fue efectivo y revisar si tienen detalles particulares sobre los otros que el modelo si pudo predecir correctamente y tratar de encontrar una relación sobre los mismos.

En caso de que no se pueda localizar un patrón en los datos antes mencionados, se debe regresar al paso 3.2.3 y volver a realizar el procesamiento de datos porque existe la probabilidad de que no se haya realizado correctamente en un inicio.

3.5. Implementación y validación.

Una vez que el modelo haya sido probado y aceptado, se debe de implementar en el ambiente para el cual fue diseñado, posteriormente probar con datos reales de la empresa y validar sus resultados para comprobar si el mismo es adecuado.

3.5.1. Implementación del modelo resultante de minería de datos

El modelo resultante debe ser implementado en el entorno real de la empresa y ser utilizado, ingresando nuevos datos los cuales deben arrojar una predicción basada en el modelo entrenado con el historial de datos brindados.

Para implementar el modelo se puede desarrollar una aplicación de software que se incorpore a los sistemas utilizados por la organización. Además esta aplicación puede arrojar los resultados de predicción en tiempo real cada vez que un nuevo candidato sea registrado.

Una vez que el sistema empiece a contar con registros reales se va a poder proceder con la medición de los resultados.

3.5.2. Medición de los resultados

Una vez que el modelo ha sido implementado, se debe de probar con datos reales los cuales deben ser registrados durante un tiempo establecido. Cabe destacar que los resultados no son definitivos, sino que son una predicción debido a la tendencia del historial de los datos que se tienen previamente.

3.5.3. Evaluación de los resultados

Cuando se tenga una cantidad considerable de resultados arrojados por el modelo, se deben evaluar para corroborar si el modelo realmente es adecuado para la necesidad de la empresa. De no ser así, se debe regresar a la primera etapa y volver a llevar todo el proceso completo dado a que existe la posibilidad de que alguno de los pasos no se haya realizado de manera correcta o eficiente.

Es importante aclarar que, debido a que es un modelo de aprendizaje automatizado, conforme nuevos datos sean ingresados, existe la posibilidad de que el resultado esperado cambie. Entonces se recomienda una revisión periódica para corroborar si el modelo sigue teniendo el mismo comportamiento, de no ser así se debe regresar a la primera etapa para definir de nueva cuenta las variables.

4. IMPLEMENTACIÓN

En este capítulo se presenta el desarrollo y la implementación del modelo propuesto en el capítulo previo, la cual fue aplicada en la empresa de desarrollo de software Nearsoft. Además, incluimos un apartado para los hallazgos sobre el los resultados de los candidatos al momento de contestar el examen técnico.

A continuación se detallan las actividades que se realizaron en cada una de las fases que componen el modelo, siendo más específicos para la organización antes mencionada.

4.1. Definición de la problemática y estructuración del objetivo.

Con la finalidad de entender y definir el problema correctamente para posteriormente establecer cual técnica de minería de datos sería la ideal para aplicar en est caso, se llevaron a cabo varias reuniones con la parte interesada con el fin de clarificar la situación en la que vivía la empresa en ese momento.

4.1.1. Definir el problema y las restricciones

En un par de reuniones concretadas para conocer la problemática actual, los interesados en la empresa mostraron su preocupación por la cantidad de candidatos potenciales que han perdido en el transcurso del proceso de reclutamiento, por lo que quisieran identificarlos de manera rápida al inicio del proceso para poderlos atender de manera prioritaria durante el proceso de evaluación y reclutamiento.

El problema presentado es una cuestión tanto de tiempo como de recursos, pues el tiempo invertido en ellos al final se desperdicia y la empresa deja de ganar dinero al no tener a dichos empleados dentro de su flotilla.

También se tocaron temas como restricciones, debido a que el problema puede abordarse de diferentes maneras y existen muchos temas que se pueden tocar. Se decidió que únicamente se va a trabajar con el examen lógico que se les aplica

en la segunda etapa para poder determinar hasta cuál etapa pudiera llegar el candidato en cuestión. También se acordó que únicamente se va a trabajar en la predicción de los candidatos con mayor potencial, pues mencionaron la posibilidad de poder conocer cuáles de los candidatos abandonan la empresa después de un determinado tiempo pero este tema está fuera del foco de investigación.

4.1.2. Definir el objetivo de la minería de datos

Después de un par de reuniones para hablar sobre la problemática y las limitaciones de la investigación, se tuvo una reunión para explicarle al departamento encargado la herramienta que se puede utilizar y la que es más conveniente para este caso, la cual es la minería de datos.

Para el caso particular de esta investigación, se opta por minería de datos dada la problemática planteada que indica la necesidad de predecir resultados de eventos dado un conjunto de datos.

Se tuvo una reunión específica para hablar sobre el objetivo que tiene esta herramienta y cuáles son las ventajas de implementar las técnicas en un conjunto de datos. Se explicaron muchos casos de éxito donde estas técnicas se han implementado y los resultados a los que se han llegado.

4.1.3. Definir los atributos y variables

Debido a que, como se mencionó en la literatura, es importante elegir correctamente los atributos que van a servir para poder trabajar. Se acordó con la empresa que no se van a considerar los aspectos sociodemográficos de los candidatos para minar los datos y hacer la predicción, pues se pretende evitar un sesgo discriminatorio por cuestiones distintas a la capacidad y habilidades necesarias para realizar las funciones propias del puesto. Por lo tanto, los datos a utilizar van a ser únicamente las calificaciones obtenidas dentro de los proceso de evaluación.

4.1.4. Seleccionar técnicas de minería de datos

Dado el tipo de datos con los que se cuenta, se decidió trabajar con Naive Bayes, ya que es una técnica de minería de datos muy utilizada y eficiente. Este clasificador asume que cada variable es independiente de las demás y en este caso nosotros asumimos que la respuesta de una pregunta en alguna de las evaluaciones no tiene influencia sobre las otras, de modo que el enfoque embona en la interpretación que le damos a la problemática.

4.2. Recolección y preparación de los datos.

Con la finalidad de poder tener datos fiables para poderlos someter a técnicas de minería de datos, fueron necesarias una serie de actividades previas para trabajar con ellos, entre las cuales se encuentran la transformación, revisión, reducción, y enriquecimiento de datos, explicados en el capitulo anterior.

4.2.1. Recopilación de los datos

La misma empresa es quien se ha encargado de recopilar los datos, adquiridos a lo largo de los últimos cuatro años.

Los datos fueron brindados por la empresa en un archivo de base de datos de MySQL, el cual fue montado en una computadora para poder acceder a ellos y empezar a analizarlos, conocer la relación entre las tablas y entender cómo es que la empresa distribuye la información sobre su proceso de reclutamiento.

4.2.2. Revisar la distribución de los datos

Una vez que se tuvo acceso a los datos, se empezaron a analizar las tablas por separado, se encontraron las siguientes tablas:

A. Rol de los candidatos
B. Catálogo de habilidades
C. Respuestas de examen
D. Preguntas de examen
E. Exámenes

F. Candidatos
G. Habilidades de candidatos
H. Respuestas de exámenes de candidatos

Cada una de las tablas contiene datos relevantes para esta investigación, en especial la tabla de candidatos, pues contiene toda la información personal de cada uno de ellos. Desde los datos sociodemográficos hasta los resultados de cada una de sus etapas en el proceso de reclutamiento.

4.2.3. Procesamiento de datos

Antes de trabajar con los datos era necesario un pre-procesamiento en el se eliminarán datos redundantes o innecesarios y crearán relaciones, en caso de no existir, sobre la información existente del conjunto de muestras de entrenamiento.

En este caso el criterio para incluir una muestra en el set de entrenamiento de nuestro modelo consistió en tener el registro de al menos 50% de las respuestas requeridas en la evaluación. Dada la naturaleza del problema nos era imposible aplicar una técnica para calcular el valor de los datos faltantes, pues no hay una relación entre ellos ni una continuidad en el tiempo.

Nuestra información no fue enriquecida con nuevos atributos producto del análisis de otras variables o de una relación entre los datos, pues no lo consideramos necesario.

Debido a que los candidatos, exámenes y respuestas se encontraban en tablas separadas, se procedió a crear una tabla con un concentrado el cual incluye el indicador único para cada candidato, un indicador único para cada examen, las 20 preguntas asociadas al examen lógico, las 4 etapas de entrevista clave (examen lógico, examen de inglés, examen de perfil y examen de programación conjunta).

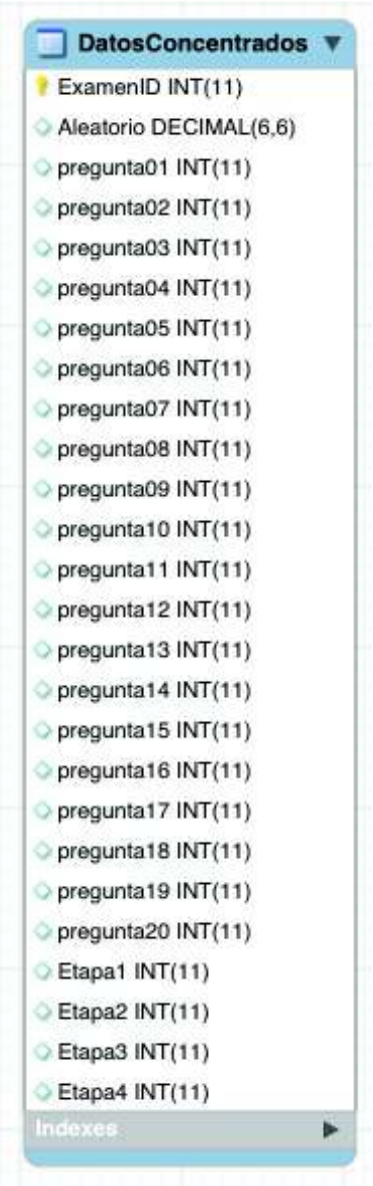

Figura 4.1 Estructura de la tabla de registros conteniendo las respuestas al examen técnico y los resultados de las etapas 1 a la 4, que fueron utilizados para el entrenamiento del modelo de predicción.

La figura 4.1 muestra la estructura de la tabla que contiene la totalidad de datos útiles para entrenar el modelo de predicción. Cada renglón de la tabla representa un registro diferente para cada uno de los candidatos, conteniendo sus respuestas al examen técnico y su desempeño en las etapas de evaluación desde la uno hasta la cuatro.

Una vez que se obtuvo una sola tabla con el concentrado de cada uno de los registros que son necesarios para empezar el análisis se decidió etiquetar los datos, esto es, asignarles un valor numérico decimal aleatorio a cada registro que van desde 0 hasta 1; esto con el fin de separarlos con una proporción definida para crear un conjunto de entrenamiento para el modelo y un conjunto de pruebas. La proporción se realizó con 90% de los datos para entrenamiento y 10% para pruebas, cada uno de estos datos fueron tomados al azar.

En nuestra solución no hubo necesidad de aplicar una técnica de reducción de datos, pues el número de atributos utilizados era perfectamente manejable.

4.3. Construcción del modelo de minería de datos

Una vez obtenidos datos completamente útiles, fueron sometidos a técnicas de minería de datos para su procesamiento y de esta manera poder construir un modelo el cual nos sirvió para probar en la empresa.

4.3.1. Implementación de técnicas de minería de datos

En este trabajo se utilizaron algoritmos de aprendizaje automatizado con el fin de extraer conocimiento de un conjunto de observaciones en los resultados de las evaluaciones a las pruebas de los candidatos con el propósito de seleccionar cuales son los que tienen mayor probabilidad de ser contratados en una etapa final dado el comportamiento de las respuestas de su primer evaluación, evitando el error humano o la parcialidad debido a los prejuicios de los perfiles**.**

La primer técnica de minería de datos con la que se empezó a trabajar los datos es Naïve Bayes, la cual fue implementada utilizando la herramienta Matlab y siguiente los pasos acontinuación descritos:

- Se construyó un conjunto de datos para entrenamiento, tomando el 90% de los datos de forma aleatoria.
- Se separaron los datos en dos clases: positivos y negativos. Los cuales indicaban el éxito o fracaso de una determinada etapa del proceso para cada candidato.

- Para cada una de las clases, se calculó la probabilidad matemática de que un valor o respuesta existiera en cada uno de los reactivos.
- Se realizó una prueba de efectividad con el conjunto del 10% de los datos restantes en el primer paso y se calculó la probabilidad de que cada registro perteneciera a cada clase.
- El criterio para medir la efectividad se dio por el porcentaje de aciertos en el paso anterior.

Después de mucho análisis y proceso de los datos dados, se produjo un modelo resultante el cual se necesitó someterse a pruebas.

4.4. Análisis de evaluación del modelo.

Con la finalidad de determinar si el modelo resultante de minería de datos es fiable, se sometió a pruebas con datos los cuales fueron reservados desde un inicio con este fin, los cuales se conocía de previa mano su comportamiento final. De esta manera pudimos evaluar si el modelo podía predecir su resultado real.

4.4.1. Someter el modelo a pruebas

Los modelos se sometieron a pruebas para validar su efectividad. Un criterio para medir la precisión de un modelo probabilístico de predicción como el que aquí implementamos, es la comparación de su exactitud contra la probabilidad de acierto al azar. En este caso, existen dos posibles resultados para cada etapa: pasar o no pasar. Si tomamos ambas opciones con la misma probabilidad, podríamos decir que el elegir un resultado de manera aleatoria sería de un 50% (como el de lanzar una moneda al aire), sin embargo, de acuerdo al historial de la empresa, las probabilidades de pasar la primera etapa es del 28%. En nuestra implementación, los resultados son bastante aceptables para la primera etapa y la precisión disminuye conforme avanzan las etapas. Este comportamiento es el esperado en una solución como la que aquí presentamos, pues los pronósticos más próximos son sobre los que se tiene una mayor cantidad de información.

Después de aplicar Naïve Bayes a los datos y generar un modelo, se procedió a utilizar el set de datos de prueba para validar la fiabilidad de dicho modelo.

El resultado de las pruebas nos arrojaron los siguientes resultados:

A. Para los candidatos que pasaron las etapas, el porcentaje de exactitud es:
    1. Etapa 1 = 100%
    2. Etapa 2 = 85%
    3. Etapa 3 = 80%
    4. Etapa 4 = 78%
B. Para los candidatos que no pasaron las etapas, el porcentaje de exactitud es:
    1. Etapa 1 = 83%
    2. Etapa 2 = 60%
    3. Etapa 3 = 50%
    4. Etapa 4 = 40%

Como se puede observar, el modelo es más preciso sobre los candidatos que pasan cada una de las etapas que sobre los candidatos que no pasan una determinada fase.

Como parte de un análisis posterior para entender los motivos por los que el modelo tiene una menor precisión al predecir a los candidatos que no pasarían a las siguientes etapas, se procedió a revisar cada uno de los registros que el modelo predijo deberían pasar una determinada etapa y los reclutadores determinaron que no era apto para continuar con el proceso de selección.

El resultado del análisis de cada uno de estos registros arrojó, en su mayoría, información importante que está fuera de nuestro control, ya que va más relacionada al perfil cultural con el cual cuentan los candidatos a la habilidad técnica del mismo.

Dentro de dicho resultado, se encontraron comentarios los cuales se concentran en la siguiente lista:

1. El candidato no sabe trabajar en equipo.
2. La actitud del candidato no es buena.
3. El candidato asumió muchas cosas.
4. No realizó las preguntas suficientes sobre las indicaciones.
5. La comunicación del candidato no es buena.
6. No tiene actitud de líder.
7. El candidato ha tenido problemas laborales con compañeros en otros empleos.
8. No toma buenas decisiones.
9. No es abierto a sugerencias o críticas.
10. Tiene actitud defensiva.
11. No escucha.
12. No sabe delegar.
13. Se limita a realizar sus tareas.

Al momento de profundizar un poco más dentro de los comentarios, se encontró que la mayoría de ellos van acompañados con buenas críticas sobre el desempeño técnico de los candidatos. Como se puede observar en la lista, todos estos comentarios son contrarios a la lista de valores que la empresa busca que las personas tengan, por lo que el desempeño técnico pesa menos al momento de evaluar.

Ya que es difícil poder predecir la cultura de cada uno de los candidatos, y la evaluación llega a ser subjetiva en función del entrevistador, existe una serie de preguntas previamente formuladas para poder deducir si el candidato cumple con cada uno de esos valores.

Dado a que la investigación no toma en cuenta los valores culturales de los candidatos, se puede proseguir con la implementación del modelo en la empresa ya que a pesar de que la predicción no fue 100% fiable, eso se debió a los factores que están fuera de la limitación los cuales son los aspectos culturales de los individuos.

4.5. Implementación y validación.

Una vez que comprobamos que el modelo funcionó de manera correcta con los datos de prueba, se sometió a nuevamente pruebas con información nueva que en ese momento ya se tenía en la compañía sobre candidatos y de esta manera pudimos determinar si el modelo era correcto.

4.5.1. Implementación del modelo resultante de minería de datos

Dentro de la empresa, se contactó tanto a la persona que funge como vínculo directo y supervisor del proyecto a implementar como al vicepresidente de operaciones y al equipo de reclutamiento. Lo anterior con el fin de mostrar el modelo propuesto y los índices de fiabilidad de los resultados con el conjunto de datos de pruebas.

Una vez que fue presentado el modelo, se procedió a asignar a una persona como asistente para la codificación digital del modelo con el fin de plasmarlo en un sistema informático, el cual sirva de apoyo al equipo de reclutamiento a la hora de trabajar con los candidatos.

Después de varios días de trabajo, se logró tener un sistema web el cual está incorporado al sistema que actualmente se usa. Esto ayudó a poder empezar a evaluar datos reales de los candidatos que van empezando su proceso.

A la fecha, se cuenta con 70 registros de candidatos que han pasado por el modelo de minería de datos; todos estos han iniciado su proceso durante el año 2017.

4.5.2. Medición de los resultados

Para la medición de resultados es necesario que los candidatos realicen el examen lógico de 20 preguntas. Una vez teniendo sus respuestas se va a poder determinar si cada uno de ellos tiene o no probabilidad de pasar una etapa determinada.

Los 70 registros reales que se han evaluado al primer bimestre del año 2017 cuentan con su predicción y algunos de ellos ya han sido evaluados en por lo menos dos etapas.

Cada vez que un candidato responde el cuestionario inicial con las 20 preguntas al inicio de su proceso, el reclutador a cargo recibe un correo electrónico de confirmación con el puntaje del examen y además de esto, información sobre si es probable que pase cada una de las etapas siguientes. Esto con el fin de conocer de ante mano la predicción de potencial de los involucrados.

Dentro de los registros medidos, se han obtenido las siguientes predicciones:

A. 29/70 tienen probabilidad de pasar la primer etapa.
B. 37/70 tienen probabilidad de pasar la segunda etapa.
C. 36/70 tienen probabilidad de pasar la tercera etapa.
D. 46/70 tienen probabilidad de pasar la cuarta etapa.

Es importante mencionar que para poder aplicar para una etapa, es necesario haber aprobado la anterior o el proceso del candidato se va a detener en ese momento y se calificará como rechazado. Habiendo aclarado lo anterior, hay ejemplos de candidatos que no es probable que aprueben la etapa 1, sin embargo su patrón indica que es probable que pasen tanto la etapa 2, 3 y 4, pudiendo llegar a ser contratados al final debido al potencial predicho.

Este comportamiento se explicó y discutió con la empresa, la cual prestó atención especial en él.

4.5.3. Evaluación de los resultados

Debido a que no existe una cantidad significante de resultados a medir, la evaluación de los resultados al día de hoy es superficial. Sin embargo se puede destacar la siguiente información:

Predicción de éxito	29
Predicción de fracaso	41
Aplicaciones realizadas	70
Fiabilidad	88.57%

Tabla 4.1 Evaluación de etapa 1.

La Tabla 4.1 muestra cómo el modelo predijo que 29 candidatos tiene probabilidad de aprobar la etapa 1, mientras que 41 no tenían probabilidad de pasarla. El 100% de los candidatos han llevado a cabo la primer evaluación y el porcentaje de fiabilidad fue del 88.57%, es decir 21 candidatos realmente pasaron la etapa 1.

Predicción de éxito	37
Predicción de fracaso	33
Aplicaciones realizadas	13
Fiabilidad	53.84%

Tabla 4.2 Evaluación de etapa 2.

La Tabla 4.2 muestra como el modelo predijo que 37 candidatos tiene probabilidad de aprobar la etapa 2, mientras que 33 no tenían probabilidad de pasarla. Sin embargo, solo 21 de los candidatos acreditaron la etapa 1 y dentro de ellos, solo 13 han presentado la segunda evaluación. Para esos 13 candidatos, el modelo acertó en un 53.84% de las veces.

Para la etapa 3, el modelo predijo que 36 candidatos tiene probabilidad de aprobar la etapa 3, mientras que 37 no tenían probabilidad de pasarla. Sin embargo, solo 10 de los candidatos aprobaron la etapa 2 dentro de ellos, solo 1 ha presentado la tercera evaluación. Dentro de esos 13 candidatos, el modelo acertó en un 53.84% de las veces. El modelo predijo que el candidato que presentó la tercera evaluación debía de aprobarla y así fue, por lo tanto la confiabilidad para esta etapa es del 100%.

Por último, para la etapa 4 el modelo predijo que 46 candidatos tiene probabilidad de pasar la etapa 4, mientras que 24 no tenían probabilidad de pasarla. Al día de hoy ninguno de los 70 candidatos ha podido presentar le etapa 4 por lo que no podemos determinar un porcentaje de fiabilidad.

Se analizaron particularmente los casos en los que nuestro modelo predijo éxito para las etapas 2, 3 y 4 que son en las cuales el factor humano interviene (entrevistadores) para determinar si existe un patrón en esas personas el cual pudiera causar el error, con el fin de descartar alguna discriminación sociodemográfica.

El 95% de los candidatos que no aprobaron una etapa son hombres, de los cuales el 80% reside en la ciudad de Hermosillo, Sonora. Estos datos nos indican que no es probable que exista una discriminación, ya que como se conoce de ante mano la mayoría de los candidatos son hombres (83%) y la empresa reside en Hermosillo.

Como se puede observar, debido a la poca cantidad de datos que se han introducido al modelo, no es posible al día de hoy terminar de evaluar ya que son pocos los candidatos que han presentado su evaluación en las etapas posteriores a la primera.

4.6. Ventajas de la Minería de Datos sobre la Heurística

Una de las soluciones más comunes a problemas como el que aquí planteamos, es la simple aplicación de filtros para seleccionar a los candidatos que hayan resultado con un puntaje por encima del mínimo requerido para pasar a la siguiente etapa. Sin embargo, en el análisis de exploración previo a la implementación se encontraron patrones en las respuestas proporcionadas por candidatos calificados que no correspondían a las correctas. Esto quiere decir que ciertas preguntas eran interpretadas de manera distinta por al menos 2 grupos de candidatos, de modo que las respuestas más frecuentes no eran las correctas.

Asumiendo que el motivo de este comportamiento pudiera ser un error en la redacción, el planteamiento o la misma interpretación del candidato, no se puede emitir un juicio en la falta del conocimiento o en la habilidad que pretende evaluar ese reactivo, sino una falla en el proceso o examen técnico, de modo que un candidato competente que conteste todos los reactivos que presentan dicho

problema con la respuesta alternativa sería descalificado sin darle oportunidad a pasar a la siguiente fase.

La tabla 4.3 muestra el concentrado de respuestas tal como se observa a continuación:

		Respuestas								
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Preguntas	**Q01**	1438	146	475	4274	79				
	Q02	1439	4359	266	348					
	Q03	1463	27	90	194	52	152	93	4115	226
	Q04	1466	425	3311	314	274	622			
	Q05	1481	428	145	1977	566	16	1799		
	Q06	1486	378	293	4038	217				
	Q07	1480	45	53	4688	146				
	Q08	1489	102	391	160	3903	367			
	Q09	1499	175	106	31	416	191	3994		
	Q10	1509	437	268	1065	3133				
	Q11	1542	963	816	520	2446	125			
	Q12	1503	28	60	118	23	14	1124	12	3530
	Q13	1508	207	3114	4	107	1277	195		
	Q14	1521	1361	180	56	1263	19	2012		
	Q15	1547	729	2617	167	145	54	39	13	1101
	Q16	1574	154	3532	244	49	42	817		
	Q17	1575	322	4319	156	40				
	Q18	1639	197	698	64	304	149	1528	1833	
	Q19	1672	389	1021	334	32	38	2926		
	Q20	1801	50	2030	360	108	185	17	1861	

Tabla 4.3 Frecuencias de respuestas en el éxamen de lógica

En la Tabla 4.3 Frecuencias de respuestas en el éxamen de lógica3 se muestra el concentrado de las respuestas dadas a las 20 preguntas del examen lógico, nótese que no todas las preguntas contienen la misma cantidad de opciones disponibles como respuesta, los campos vacíos representan opciones no disponibles para una determinada pregunta.

La tabla 4.4 muestra el resumen de las preguntas con la respuesta correcta y la respuesta con mayor incidencia dentro el histórico de los candidatos.

		Respuesta	
		Mas elegida	**Correcta**
Preguntas	**Q01**	3	3
	Q02	1	1
	Q03	7	7
	Q04	2	2
	Q05	3	3
	Q06	3	3
	Q07	3	3
	Q08	4	4
	Q09	6	6
	Q10	4	4
	Q11	4	4
	Q12	8	8
	Q13	2	5
	Q14	6	1
	Q15	2	2
	Q16	2	2
	Q17	2	2
	Q18	7	6
	Q19	6	6
	Q20	2	7

Tabla 4.4 Relacion entre respuesta correcta y respuesta con mayor incidencia en el examen lógico

Como se puede observar en la tabla 4.4, en la mayoría de los casos la respuesta correcta es la que tiene mayor frecuencia, sin embargo existen 4 preguntas las cuales se respondieron de manera errónea en una mayor cantidad de veces, estas últimas preguntas fue donde pusimos especial atención.

Para poder entender porque sucede el fenómeno descrito anteriormente, fue necesario analizar cada una de los reactivos del examen lógico, el tipo y complejidad de los mismos.

Dentro del análisis se encontraron detalles importantes, los cuales se localizan en la manera en la que están descritas las instrucciones de cada pregunta. Se pudo observar que cada pregunta contiene una instrucción específica sobre como determinar la respuesta; sin embargo, las 4 preguntas que se mencionaron anteriormente solo pueden ser respondidas de manera correcta dada una concatenación de instrucciones, es decir, hay que poner atención en instrucciones de reactivos anteriores y agregarlos a la pregunta en cuestión.

La mayoría de las personas no siguen las instrucciones dadas, lo que los lleva a elegir una respuesta equivocada con una frecuencia bastante alta en el examen lógico. Esto puede deberse a la falta de atención o al estrés del evento.

La implementación del modelo permitió la detección de este problema, pues proporcionó casos que cumplían con el comportamiento de un candidato que sí pasó a la etapa final, sin embargo no pudo continuar en el proceso dado que contestó las preguntas mencionadas anteriormente con la segunda respuesta con mayor frecuencia, sin embargo, esto no significa que el candidato no cumple con los requerimientos técnicos, sino que la formulación de la pregunta, el ambiente, o hasta el mismo nerviosismo impida que tenga un desempeño óptimo al momento de la evaluación. Este enfoque permite corregir el método de evaluación y los reactivos que en ella intervienen. La aplicación de estos modelos de selección de personal pueden proporcionar una ventaja sobre los convencionales, pues son capaces de aprender y detectar los casos donde el escenario no favorezca al candidato, aun cuando tenga el conocimiento y habilidades requeridas.

En este estudio sólo se tomaron en cuenta los conocimientos técnicos y el desempeño de los aplicantes durante la evaluación, sin embargo, el aspecto cultural (soft skills) juega un papel determinante en la decisión final.

5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

En la presente investigación, se buscó desarrollar e implementar un modelo para predecir el éxito de los candidatos que inician el proceso de reclutamiento en una empresa de desarrollo de software.

A continuación, se describen las conclusiones, recomendaciones y trabajos futuros relacionados con el trabajo desarrollado.

5.1. Conclusiones.

En un entorno donde existe la necesidad de contratar al personal más competente, es altamente probable que el proceso sea lento y tedioso, pues se busca tener los mejores filtros para lograrlo.

Por tal motivo, se desarrolló un modelo con la capacidad de ser aplicado a entornos donde la contratación de personal capacitado sea una prioridad, así como la posibilidad de ser adaptable a otros procesos de reclutamiento.

El modelo fue propuesto, evaluado e implementado en la organización, el cual fue capaz de predecir el número de etapas que puede aprobar un candidato, de modo que fue posible detectarlos oportunamente para darles seguimiento de manera inmediata.

Para poder implementar el modelo de la manera mas eficiente, se buscó utilizar herramientas tecnológicas que ayudaran a facilitar los procesos de análisis de y evaluación de datos, así como la elección de la técnica de minería de datos que mayor fiabilidad nos brinde. La elección de la técnica correcta de minería de datos fue primordial para lograr el éxito del proyecto.

El equipo de reclutamiento y la selección de personal juegan un papel que se ve reflejado directamente en la calidad y desempeño de los trabajadores. Actualmente estos procesos son reforzados por técnicas como la ayuda para la toma de decisiones y tecnologías de la información. Un modelo de clasificación para la detección de los mejores candidatos en sus primeras evaluaciones puede

revolucionar los procedimientos de reclutamiento haciéndolos más dinámicos y eficientes valiéndose del historial de contrataciones y la información actual de los aspirantes.

En áreas de desarrollo tecnológico donde predomina un estereotipo del empleado ideal, resulta difícil luchar contra el sesgo en las decisiones que determinan una contratación; en estos casos, la automatización de los procesos de selección con apego estricto a los requerimientos técnicos ayudan a solventar el problema de la discriminación. Sin embargo, hay factores que requerirán de intervención humana, como la evaluación de la compatibilidad del equipo con el que ha de integrarse el candidato.

La minería de datos puede jugar un papel importante para ayudar a los procesos de selección al predecir cuáles candidatos siguen un patrón de éxito y de esta manera identificar a estos elementos claves para darles prioridad sobre otros reduciendo el riesgo de perderlos durante el proceso, mejorando la calidad de los empleados que redundará en mejores productos en la compañía, un aumento en las ganancias y reducción de costos a largo plazo.

5.2. Recomendaciones.

Este trabajo de tesis fue desarrollado con la finalidad de generar un modelo que pudiera implementarse en cualquier empresa de desarrollo de software con el mínimo de adecuaciones, por lo que es importante realizar un buen diagnóstico y evaluación del proceso de reclutamiento durante la fase inicial, así como la elección correcta en la técnica de minería de datos a implementar al momento de aplicar el modelo.

La principal recomendación consiste en la aplicación de un cuestionario que obtenga información cultural relevante para definir la personalidad del candidato y determinar si cubre los requisitos para trabajar en equipo, si el candidato es proactivo y tiene aptitudes de liderazgo, y otras que pudieran ser relevantes a la hora de la selección. Estos datos pudieran complementar el modelo desarrollado

en este trabajo para aumentar el nivel de acierto en la predicción a la hora de la selección de los mejores candidatos.

Toda adquisición de datos relacionados con aspectos de índole personal deberá tener los niveles más altos de confidencialidad y garantizar a quien se somete al proceso, que no existirá un riesgo de filtración ni un link entre sus datos y su identidad para fines ajenos a la evaluación en proceso.

5.3. Trabajos Futuros.

Un trabajo a futuro debiera incluir un test o exploración previo al proceso de reclutamiento que cubra el aspecto cultural, la cuales nos permitan obtener un par de datos extras sobre los candidatos y permita mayor precisión al momento del filtrado y predicción.

Este proyecto ha sido presentado ante la empresa y se ha tomado la decisión de formar una comisión para el desarrollo de un sistema automatizado que aplique el modelo aquí presentado y sirva de apoyo al área de recursos humanos para la selección de personal.

En la siguiente etapa de este proyecto se enfocará a probar con otros algoritmos y técnicas de minería de datos e inteligencia artificial para determinar la consistencia en los resultados. Se profundizará más en los casos rechazados por un enfoque y aceptados por otros, para encontrar los factores que más influyen en la decisión y determinar si son relevantes o no.

6. REFERENCIAS

Agarwal, N. y Liu, H. y Zhang, J., 2006. Blocking Objectionable Web Content by Leveraging Multiple Information Sources. *ACM SIGKDD Explorations Newsletter*, pp.17–26.

Badr El Din Ahmed, A. y Sayed Elaraby, I., 2014. Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*, 2(2), pp.43–47.

Beckers, A.M. y Bsat, M.Z., 2002. A DSS classification model for research in Human Resource Information Systems. *Information Systems Management*, 19(3), pp.41–50.

Bello, M. y Bello, R. y Nowé, A. y García-Lorenzo, M., 2016. Personnel Selection in a Competitive Environment. *Computación y Sistemas*, 20(2), pp.195–204.

Berry, M.J.A. y Linoff, G.S., 1999. *Mastering Data Mining - The Art and Science of Customer Relationship Management*.

Berry, M.J. y Linoff, G., 1997. *Data mining techniques: For marketing, sales and customes support*, John Wiley & Sons.

Betts, M., 2003. The Almanac: Hot Tech. *ComputerWorld*.

Borman, W.C. y Hanson, M.A. y Hedge, J.W., 1997. PERSONNEL SELECTION. *Annual Review of Psychology*, 48(1), pp.299–337. Available at: http://www.annualreviews.org/doi/10.1146/annurev.psych.48.1..

Chen, M.S. y Han, J. y Yu, P.S., 1996. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp.866–883.

Chien, C. y Chen, S. y Lin, Y., 2002. Using Bayesian network for fault location on distribution feeder. *IEEE Trrancsactions on Power Delivery*, 17(13), pp.785–793. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1022804.

Chien, C.F. y Chen, L.F., 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34(1), pp.280–290.

Chien, C.F. y Tseng, T.L.. y Peng, J.-T., 2004. Rough set theory for data mining for fault diagnosis on distribution feeder. *IEE Proceedings-Generation, Transmission and …*, 151(6), pp.689–697. Available at: http://digital-library.theiet.org/content/journals/10.1049/ip-gtd_20040098.

DeBarr, D. y Eyler-Walker, Z., 2006. Closing the gap: automated screening of tax returns to identify egregious tax shelters. *ACM SIGKDD Explorations Newsletter*, 8(1), pp.11–16.

DeNisi, A.S. y Wilson, M.S. y Biteman, J., 2014. Research and practice in HRM: A historical perspective. *Human Resource Management Review*, 24(3), pp.219–231. Available at: http://dx.doi.org/10.1016/j.hrmr.2014.03.004.

Fayyad, U. y Piatetsky-Shapiro, G. y Smyth, P., 1996. Advances in Knowledge Discovery and Data Mining. , p.560.

Freeman, E. y Melli, G., 2006. Championing of an LTV model at LTC. *ACM SIGKDD Explorations Newsletter*, 8(1), pp.27–32.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), pp.452–459. Available at: http://dx.doi.org/10.1016/j.cie.2010.09.004.

Ghani, R. y Probst, K. y Liu, Y. y Krema, M. y Fano, A., 2006. Text Mining for Product Attribute Extraction. *ACM SIGKDD Explorations Newsletter*, pp.41–58.

Giudici, P., 2003. *Statistical Methods for Business and Industry*.

Gorgoso, J.J. y González, J.G.Á. y Rojo, A., 2007. Modelling diameter distributions of Betula alba L . stands in northwest Spain with the two-parameter Weibull function. , 16(2), pp.113–123.

Goth, G., 2004. E-Volting Milestones. *IEEE Security and Privacy*, 2(1), p.14.

Grenander, U. y Miller, M.I., 1998. Computational Anatomy: an Emerging Discipline. *Quarterly of Applied Mathematics*, 56(4), pp.617–694.

Gupta, S. y V, S., 2013. Empirical Study on Selection of Team Members for Software Projects - Data Mining Approach. *International Journal of Computer Science and Informatics*, 3(2), pp.97–102. Available at: http://arxiv.org/abs/1402.2377.

Ha, S.H. y Bae, S.M. y Park, S.C., 2000. Web mining for distance education. *Proceedings of the 2000 IEEE International Conference on Management of Innovation and Technology:* , 2, pp.715–719.

Ha, S.H. y Park, S.C., 2006. Service Quality Improvement through Business Process Management based on Data Mining. *ACM SIGKDD Explorations*, 8(1), pp.49–56.

Hernandez Sampieri, R. y Fernandez Collado, C. y Baptista Lucio, P., 2014. *Metodología de la investigación* 6th ed.

Hooper, R. y Galvin, T. y Kilmer, R. y Liebowitz, J., 1998. Use of an expert system in a personnel evaluation process. In *Expert Systems with Applications*. pp. 425–432.

Jantan, H. y Hamdan, A.R. y Othman, Z.A., 2010. Human Talent Prediction in HRM using C4 . 5 Classification Algorithm. *(IJCSE) International Journal on COmputer Science and Engineering*, 2(October 2015), pp.2526–2534.

Kahraman, C., 2015. *Intelligent Techniques in Engineering Management*.

Kanchanamani, M. y Perumal, V., 2016. Performance evaluation and comparative analysis of various machine learning techniques for diagnosis of breast cancer. *Biomedical Research*, 27(3), pp.623–631.

Kanellopoulos, Y. y Domupulos, T. y Thorthis, C. y Makris, C., 2006. Mining Source Code Elements for Comprehending OO System and Evaluating Their Maintainability. *ACM SIGKDD Explorations Newsletter*, pp.33–40.

Kimball, R. y Ross, M., 2002. *The Data Warehouse Toolkit*.

Kovach, K.A. y Cathcart, C.E.J., 1999. Human Resource Information Systems (HRIS): Providing Business with Rapid Data Access, Information Exchange and Strategic Advantage. *Public Personnel Management*, 28(2), p.275.

Liao, S., 2003. Knowledge management technologies and applications—literature review from 1995 to 2002. *Expert Systems with Applications*, 25(2), pp.155–164. Available at: http://www.sciencedirect.com/science/article/pii/S0957417403000435.

Lievens, F. y Dam, K. Van y Anderson, N., 2002. Recent trends and challenges in personnel selection. *Personnel Review*, 31(5), pp.580–601.

Lin, H.-T., 2010. Personnel selection using analytic network process and fuzzy data envelopment analysis approaches. *Computers & Industrial Engineering*, 59(4), pp.937–944. Available at: http://dx.doi.org/10.1016/j.cie.2010.09.004.

Melli, G. y Zaïane, O.R. y Kitts, B., 2006. Introduction to the special issue on successful real-world data mining applications. *ACM SIGKDD Explorations Newsletter*, 8(1), pp.1–2. Available at: http://portal.acm.org/citation.cfm?doid=1147234.1147235.

Mena, J., 2011. *Data Mining Concetps and Techniques.pdf*, CRC Press.

Mishra, T. y Kumar, D. y Gupta, S., 2014. Mining students' data for prediction performance. In *International Conference on Advanced Computing and Communication Technologies, ACCT*. pp. 255–262.

Nussbaum, M. y Singer, M. y Rosas, R. y Castillo, M. y Flies, E. y Lara, R. y Sommers, R., 1999. Decision support system for conflict diagnosis in personnel selection. *Information & Management*, 36, pp.55–62.

Peng, C. y Chien, C.F., 2003. Data value development to enhance yield and maintain competitive advantage for semiconductor manufacturing. *International Journal of Service Technology and Management*, 4(4–6), pp.365–383.

Pnheiro, C. y Evsukoff, A. y Ebecken, N., 2006. Revenue Recovering with Insolvency Prevention on a Brazilian Telecom Operator. *ACM SIGKDD Explorations Newsletter*, pp.65–70.

Rao, B. y Krishnan, S. y Niculescu, R., 2006. Data Mining for Improved Cardiac Care. *ACM SIGKDD Explorations Newsletter*, pp.3–10.

Robertson, I.T. y Smith, M., 2001. Personnel selection. *Journal of Occupational and Organizational Psychology*, 74(4), pp.441–472.

Shaw, M. J. y Subramaniam, C. y Tan, G. W. y Welge, M. E., 2001. Knowledge management and data mining for marketing. *Decision Support Systems*, 31(1), pp.127–137.

Shiue, Y.R. y Su, C.T., 2003. An enhanced knowledge representation for decision tree based learning adaptive scheduling. *International Journal of Computer Integrated Manufacturing*, 16(1), pp.48–60.

Singh, P. y Thomas, C. y Sepulveda, A., 2006. Market Basket Recommendations for the HM SMB Store. *ACM SIGKDD Explorations Newsletter*, pp.57–64.

Singh Thakur, G. y Gupta, A. y Gupta, S., 2015. Data Mining for Prediction of Human Performance Capability in the Software-Industry. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 5(2), pp.1–11.

Strohmeier, S. y Piazza, F., 2015. Artificial Intelligence Techniques in Human Resource Management—A Conceptual Exploration. *Springer International Publishing Switzerland*, 87, pp.149–172. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84929094817&partnerID=tZOtx3y1.

Vapnik, V., 1995. *The Nature of Statistical Learning Theory*, New York: Springer.

Wang, J., 2003. *Data Mining: Opportunities and Challenges*,

Wei, C.P. y Chiu, I.T., 2002. Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications*, 23(2), pp.103–112. Available at: http://eprints.utas.edu.au/1235/.

Weiss, S.H. y Indurkhya, N., 1998. *Predictive Data Mining: A Practical Guide*, San Francisco, CA: Morgan Kaufmann Publishers.

Wu, C. H. y Kao, S. C. y Su, Y. Y. y Wu, C. C., 2005. Targeting customers via discovery knowledge for the insurance industry. *Expert Systems with Applications*, 29(2), pp.291–299.