



UNIVERSIDAD
DE PIURA

REPOSITORIO INSTITUCIONAL
PIRHUA

GUÍA METODOLÓGICA PARA OBTENER PATRONES DE ACCIDENTABILIDAD LABORAL USANDO DATA MINING

Alvaro Cevallos-Chacón

Piura, octubre de 2013

FACULTAD DE INGENIERÍA

Maestría en Dirección Estratégica en Tecnologías de la Información

Cevallos, A. (2013). *Guía metodológica para obtener patrones de accidentabilidad laboral usando Data Mining*. Tesis de maestría en Dirección Estratégica en Tecnologías de la Información. Universidad de Piura. Facultad de Ingeniería. Piura, Perú.



Esta obra está bajo una [licencia Creative Commons Atribución-NoComercial-SinDerivadas 2.5 Perú](#)

Repositorio institucional PIRHUA – Universidad de Piura

UNIVERSIDAD DE PIURA
FACULTAD DE INGENIERIA



“Guía metodológica para obtener patrones de accidentabilidad laboral usando Data Mining”

Tesis para optar el Grado de Master en Dirección Estratégica en Tecnologías de la
Información

Alvaro Alexssander Cevallos Chacón

Asesor: Federico Wiese López

Piura, Octubre 2013

Este trabajo va dedicado a mi madre pues
sin su inconmensurable e inmedible amor
que junto a su coraje, valentía, ternura,
comprensión y ejemplo ninguno de mis
logros y fracasos hubiesen tenido el
significado que ahora tienen, estos son
propiedad de ella, pues su amor diseñó
todo.

Y no podría dejar de recordar a ti que estás
en el cielo, pienso feliz en el momento en
que pueda volver a sentir tu abrazo paterno.
Tú eres mi prueba lógica sobre la verdadera
existencia de Dios.

Prólogo

Ciertamente el análisis de accidentabilidad actual responde a preguntas inmediatas sobre las características de los accidentes y la legislación no se muestra explícita sobre los métodos de trabajo para este fin, también es cierto que los resultados de recurrencia de accidentes demuestran que por un lado las causas determinadas por el análisis no son las correctas o que las estrategias de mejora continua no son adecuadas.

El constante cambio que enfrentan las empresas tanto en el ámbito comercial, legal, social y laboral los obliga a manejar de manera más eficiente la información con la cual disponen y hacer todo lo posible por obtenerla de la manera más económica. Los datos ya no resultan suficientes, se requiere información que genere conocimiento duradero y objetivo de la real situación que ocurre.

La minería de datos es una tecnología que precisamente trabaja datos para obtener conocimiento usando distintos algoritmos y técnicas. Esto pues resulta en una herramienta ideal para analizar patrones de comportamiento de los datos asociados a la accidentabilidad registrada a fin de entender que factores inciden en la ocurrencia de accidentes a través de patrones. Las visiones macro y personalizadas que ofrece esta tecnología permitirán al personal estratégico comprender mejor las causas para diseñar estrategias más eficaces de prevención.

El presente trabajo propone una guía metodológica para obtener patrones de accidentabilidad laboral usando tecnología de minería de datos. El documento muestra paso a paso como con el uso de la metodología KDD *Knowledge Discovery in Database*, se puede obtener conocimiento a partir de data de accidentes de trabajo de cualquier tipo de industria. Así mismo, en cada etapa del proceso se hacen comentarios y ejemplos de posibles casos que se pueden presentar en la vida real orientando su interpretación al eventual analista que utilice esta guía para realizar su propia aplicación. El documento combina los conceptos de seguridad industrial relacionados con la investigación de accidentes y toma como marco principal, la normatividad nacional vigente en materia de seguridad y salud en el trabajo. Estos conceptos son parámetros de vital importancia para la aplicación del *data mining*.

Finalmente, el presente trabajo realiza la aplicación de la metodología en el estudio de patrones de accidentabilidad de la empresa IMI del Perú SAC con el debido consentimiento de sus líneas de dirección. Los resultados obtenidos y las diferentes formas de análisis que se pueden explotar con las herramientas usadas para hacer la minería de datos, son consultados

con personal experto. Este personal forma parte de un equipo de trabajo encargado de la gestión de la seguridad y salud ocupacional de la empresa.

Un agradecimiento personal al equipo humano que participó en este proyecto, tanto asesores como personal experto y a las gerencias de IMI del Perú SAC quiénes desde el surgimiento de la idea del autor sobre este estudio, se mostraron de acuerdo y proactivos:

Gracias a la Gerencia General de IMI del Perú, Sr. Almirante Luis Illescas Castañeda quién desde la idea primitiva de este proyecto se mostró entusiasta y brindo su venia para materializarlo. Gracias a la Gerencia de Operaciones de IMI del Perú, Sr. Comandante Walter Granda Sponholz quién a pesar de su carácter impaciente pero firme, tuvo paciencia conmigo para con mis ideas y proyectos. Gracias a la Gerencia Financiera de IMI del Perú, Sr. CPC Carlos Benavides quién también dio su apoyo favorable desde la idea del proyecto. Gracias al Ing. Elvis Albán quién más que mi jefe, fue y es mi tutor técnico. Gracias al Ing. Samuel Oporto, experto en minería de datos, probablemente uno de los profesionales más capacitados en esta tecnología a nivel nacional. Desde que lo conocí a través del blog de la SPC me brindó su apoyo técnico, siempre me atendió. Él es un ejemplo de vocación y cariño por el trabajo que a uno le gusta. Gracias a mis compañeros del Departamento QHSE quiénes me dieron su soporte y aporte nutriendo el trabajo en cada etapa con sus críticas constructivas y soberbios análisis, gracias Jaime Varillas, gracias Victor Robles, gracias Luis Alzamora, gracias Ego Castillo y gracias Milagros Panta.

Resumen

El presente trabajo ha sido motivado por la necesidad de contar con herramientas que permitan estudiar y conocer mejor la dinámica de accidentabilidad de una empresa. El objetivo fundamental es presentar un aporte al personal experto y dedicado a la gestión de la seguridad industrial y salud en el trabajo, un método de trabajo para aplicar *data mining* a las bases de datos de accidentabilidad laboral de cualquier empresa a fin que con el conocimiento obtenido se pueda diseñar estrategias de prevención eficaces para reducir los accidentes laborales así como ofrecer líneas de investigación y estudio en las cuales se explote y aproveche al máximo las bondades brindadas por las técnicas y tareas de la minería de datos.

Para realizar el estudio ha sido necesario estudiar la misma tecnología de minería de datos y las experiencias registradas en ese sentido en distintos rubros a fin de utilizar las herramientas más adecuadas así como también, entender la teoría de seguridad industrial asociada a los accidentes de trabajo, dado que la metodología propuesta representa una aplicación de tecnología en una disciplina específica.

El trabajo no estaría completo si es que no se realiza una aplicación real por lo cual con autorización de la empresa IMI del Perú SAC se practicó el estudio de patrones de accidentabilidad de su base de datos. La metodología y el conocimiento obtenido fue presentado al personal de esta empresa para su revisión y evaluación obteniéndose resultados satisfactorios y la posibilidad de usar la metodología como referencia para aplicación en otros rubros diferentes tales como el control de combustible.

ÍNDICE

INTRODUCCIÓN	1
Capítulo 1 Marco Teórico.....	3
1.1. Seguridad y salud en el trabajo	3
1.1.1. Seguridad e higiene industrial.....	3
1.1.2. Peligro, riesgo y consecuencia.	4
1.2. Accidentes de trabajo	4
1.2.1. Definición de los accidentes de trabajo	4
1.2.2. Situación de accidentabilidad laboral en el Perú	5
1.2.3. Clasificación de los accidentes de trabajo.	6
1.2.4. Anuario de Accidentabilidad del Mintra	6
1.2.5. La investigación de accidentes.....	8
1.2.6. Legislación en materia de accidentabilidad laboral	10
1.3. Minería de Datos.....	11
1.3.1. Introducción a la Minería de Datos.....	11
1.3.2. <i>Business intelligence</i>	11
1.3.3. Proceso de descubrimiento de conocimiento.....	12
1.3.4. Técnicas de Minería de Datos.....	13
Capítulo 2 Guía metodológica para obtener patrones de accidentabilidad laboral basado en <i>data mining</i>	15
2.1. Diseño de una estructura de datos de accidentabilidad laboral.....	16
2.1.1. Procedimiento de investigación de accidentes.....	17
2.1.2. Gestión de la data de accidentabilidad	18
2.1.2.1. Modelo de datos según lo requerido por la legislación peruana.	19
2.1.2.2. Datos relevantes.	20
2.1.2.3. Análisis descriptivo de datos en relación al modelo de gestión de la empresa.	26
2.2. Obtención de conocimiento a partir de los datos de accidentabilidad laboral.....	27
2.2.1. Aplicación del procedimiento estándar KDD	28
2.2.2.1. Selección, Exploración, limpieza y transformación.	29
2.2.2.2. Planteamiento del modelo de data mining.....	33
2.2.2.3. Evaluación e interpretación de resultados	36
2.3. Aplicación del conocimiento obtenido al programa de Seguridad y Salud en el Trabajo	38
Capítulo 3 Conclusiones y recomendaciones	41
Bibliografía	43
Anexos	

Anexo A - Listado de tablas y gráficos	45
Anexo B – Cuadro detallado de la ley SST respecto de la gestión de accidentes....	46
Anexo C – Aplicación real del modelo de obtención de conocimiento.....	47
1. La Empresa.....	47
2. Solución propuesta al problema específico.....	47
2.1. Fuentes de información para el análisis.....	47
2.2. Atributos de interés.....	48
2.3. Atributos no considerados.....	49
2.4. Depuración de registros	50
2.5. Limpieza y transformación.....	56
2.6. <i>Dataset</i> final.....	57
3. Desarrollo del experimento.....	58
3.1 <i>Clustering</i> (agrupamiento).....	58
3.2 Implementación del algoritmo de inducción C4.5 para la tarea de clasificación de los <i>clusters</i>	69
3.3 Árbol de Clasificación.....	71
4. Recomendaciones y conclusiones.....	79
Anexo C.1 - Listado de tablas y gráficos.....	80
Anexo C.2 - Listado de cargos o puestos de trabajo definidos según organigrama de IMI (instancias posibles de la variable cargo).....	81
Anexo C.3 - Listado de locaciones o lugares en donde se realizan las operaciones de IMI (instancias posibles de la variable Lugar).....	82

Introducción

El proyecto surgió con la idea de aprovechar las herramientas y técnicas que tiene la minería de datos a fin de obtener patrones o características de los accidentes de trabajo que ocurren en IMI del Perú SAC esto en primer lugar tiene el objetivo de robustecer el análisis estadístico que normalmente se aplica y por otro lado, obtener conocimiento que sirva como *input* para el planteamiento de medidas preventivas eficaces para minimizar los accidentes laborales. A partir de allí surgió la necesidad de formalizar y documentar una metodología específica para el estudio de este tipo de base de datos cuya aplicación pueda ser de alcance de cualquier tipo de empresa y en cualquier tipo de industria.

El trabajo está compuesto por tres capítulos. El primer explica el estado del arte en materia de seguridad industrial y minería de datos. Se exploran conceptos básicos en materia de accidentabilidad laboral y la situación general que según registros oficiales, se presenta en el Perú. También se explican conceptos teóricos y generales sobre la minería de datos, principalmente enfocándose en sus técnicas y tareas.

En el segundo capítulo se procede a exponer la metodología de trabajo en sí. Se usa como referencia dos parámetros básicos: la metodología KDD y la legislación peruana en materia de seguridad y salud en el trabajo. Así se obtiene una lista de pasos y recomendaciones a considerar para realizar un estudio de patrones de accidentabilidad. Un punto importante es la formulación de casos reales que el eventual analista lector de este estudio debe tomar en cuenta en su aplicación particular.

En el tercer capítulo se describen las recomendaciones y conclusiones del trabajo enfocándose en su carácter general y la posibilidad de extender las líneas de investigación para el estudio de accidentes laborales y medidas de prevención eficaces.

Finalmente y ya como se ha comentado anteriormente, el trabajo estaría incompleto si es que no se formula una aplicación real, pues eso también es importante para que el lector de este trabajo tenga mejores referencias de como plantear su modelo y sobre todo como lidiar con los tecnicismos propios del *data mining*. Se aplica el modelo a la base de datos de accidentes de IMI del Perú SAC y los resultados son consultados con personal experto quienes asesoran sobre los distintos enfoques que se pueden aportar al análisis. El modelo utiliza dos tareas de minería de datos: el agrupamiento y la clasificación. En el primer caso se utiliza el algoritmo *K means* y en el segundo, el algoritmo de inducción C4.5.

Capítulo 1

Marco teórico

1.1. Seguridad y salud en el trabajo

1.1.1. Seguridad e higiene industrial

El concepto de Seguridad e Higiene el Trabajo no es un concepto fijo, sino que por el contrario, ha sido objeto de numerosas definiciones que, con el tiempo, han ido evolucionando de la misma forma que se han producido cambios en las condiciones y circunstancias en que el trabajo se desarrollaba¹. Actualmente este concepto ha evolucionado en una disciplina que usa diversas técnicas para analizar la forma más segura para realizar un trabajo; es decir, evitar accidentes o lesiones sobre las personas que lo realizan y evitar enfermedades producto del trabajo al corto, mediano y/o largo plazo.

En esta misma línea de ideas se presenta la Higiene Industrial como la disciplina que se encarga de estudiar los riesgos crónicos asociados a la salud del trabajador al largo plazo. Se podría resumir ambos conceptos en la siguiente idea: la seguridad se ocupa de los efectos agudos de los riesgos, en tanto la salud trata sus efectos crónicos².

Entonces pues, Seguridad e Higiene son dos conceptos generales e interrelacionados que estudian las técnicas más adecuadas para prevenir lesiones y enfermedades en los trabajadores de cualquier tipo de industria. Este objetivo se presenta como muy ambicioso pues a la actualidad el número de lesiones o enfermedades producto del trabajo, si bien es cierto ha disminuido o se viene controlado, está muy lejos de llegar a cero y de hecho se considera como un argumento objetivo que es probablemente imposible o cuanto menos, muy difícil lograr ese valor. Es precisamente este argumento el factor que motiva el estudio continuo de las técnicas y formas que permitan tener trabajos seguros, es decir sin lesiones o enfermedades a largo plazo, motivación también del presente estudio de patrones de accidentabilidad utilizando la tecnología de minería de datos.

¹ Cortez, José. (2007). *Seguridad e higiene del trabajo. Técnicas de prevención de riesgos laborales* 9º ed. Tebar. Pag. 43.

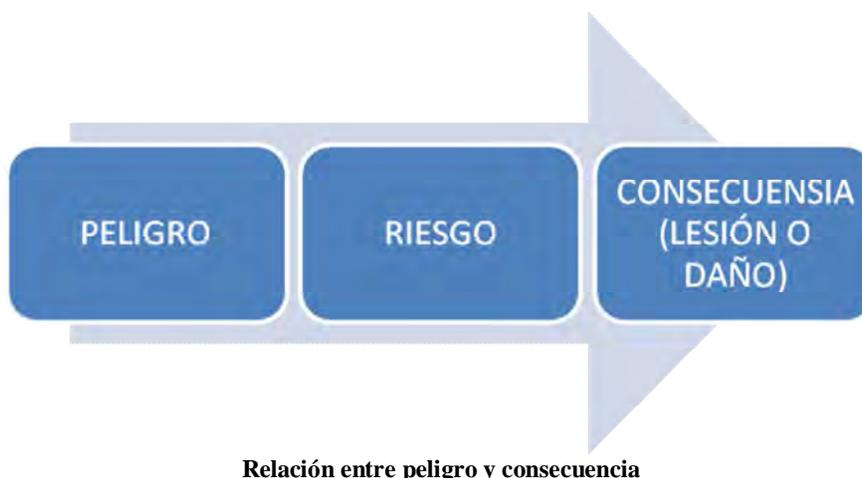
²Asfahl, C. Ray. (2000). *Seguridad industrial y salud*. 4º ed. Prentice Hall. Pag. 4.

1.1.2. Peligro, riesgo y consecuencia.

Así mismo es importante reconocer dos conceptos básicos y muy importantes que en materia de seguridad y salud en el trabajo representan el pilar fundamental de cualquier estudio: Peligro y Riesgo. A saber, el Peligro es definido como la situación o característica intrínseca de algo capaz de ocasionar daños a las personas, equipos, procesos y ambiente; mientras que el Riesgo es definido como la probabilidad de que un peligro se materialice en determinadas condiciones y genere daños a las personas, equipos y el ambiente³.

Un concepto final para complementar lo antes mencionado es la Consecuencia, que es el resultado de la materialización del riesgo; es decir, es la generación de la lesión (efecto agudo) o enfermedad ocupacional (efecto crónico) en el trabajador. A continuación se representa un esquema que relaciona estos conceptos:

Figura 1.1. **Relación entre peligro, riesgo y consecuencia.**



1.2. Accidentes de trabajo

1.2.1. Definición de los accidentes de trabajo

Los accidentes de trabajo son definidos por el Reglamento de Seguridad y Salud en el Trabajo como todo suceso repentino que sobrevenga por causa u ocasión del trabajo y que produzca en el trabajador una lesión orgánica, una perturbación funcional, una invalidez o la muerte. <...>⁴. Una forma eficaz de proponer medidas preventivas para evitar accidentes laborales es precisamente estudiarlos; sin embargo la generalización de las conclusiones de un accidente a otras situaciones conlleva a un cierto riesgo⁵; es decir que aunque el estudio puede llevarnos a conclusiones sobre las causas de los accidentes siempre estará limitado al caso específico de estudio.

³ Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Glosario de términos.

⁴ Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Glosario de términos.

⁵ Saari, Jorma. Accidentes y Gestión de la Seguridad. Enciclopedia de Salud y Seguridad en el Trabajo de la OIT. Pag. 56.2. Disponible en <http://www.insht.es/0057F17F-E555-4734-B4E9-5020B0978F88/FinalDownload/DownloadId-E81887714E078E9BA83EDE7E6D79BEA1/0057F17F-E555-4734-B4E9-5020B0978F88/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

Es particularmente importante tener claro que el estudio de patrones de accidentabilidad puede generar conclusiones erradas si es que no se realiza una interpretación adecuada de las mismas. En ese sentido, se deben escoger cuidadosamente las técnicas de minería de datos a usar pues el modelo partirá de un *dataset*, vista minableo repositorio de datos listos para ser trabajados, que contienen las características más importantes de los accidentes que se han registrado en la empresa en un intervalo de tiempo específico.

Existen cinco tipos fundamentales de análisis de accidentes⁶:

1. Análisis y determinación de los tipos de accidentes y los lugares en que se produjeron.
2. Análisis a partir del control de la incidencia de accidentes.
3. Análisis para establecer prioridades entre diferentes iniciativas que exigen un nivel elevado de medición de riesgos, lo que a su vez exige el cálculo de la frecuencia y gravedad de los accidentes.
4. Análisis para determinar cómo han ocurrido los accidentes y, sobre todo, para establecer las causas tanto directas como indirectas.
5. Análisis para dilucidar qué áreas especiales han suscitado curiosidad por alguna razón (se trata de una forma de análisis de revisión o control).

Como se ha comentado anteriormente, el hecho de generar conclusiones generales sobre el análisis de accidentes puede resultar errado de modo que la participación de un experto para en función del contexto en el cual ocurrieron los hechos, se puedan proponer interpretaciones más realistas de modo que sirvan objetivamente a la propuesta de medidas preventivas eficaces.

1.2.2. Situación de accidentabilidad laboral en el Perú

De acuerdo con las estadísticas de la Oficina Internacional del Trabajo (OIT), se producen cada año 120 millones de accidentes laborales en los lugares de trabajo de todo el mundo. De éstos, en 210 000 se registran fallecimientos. Cada día, más de 500 hombres y mujeres no regresan a sus hogares víctimas de este tipo de accidentes mortales⁷.

Las cifras descritas en el párrafo anterior resultan desmotivadoras, aunque la situación ha mejorado mucho no se ha logrado reducir drásticamente la incidencia de accidentes laborales. En el Perú la situación no es del todo diferente, el alto grado de informalidad que se vive en el país es un excelente caldo de cultivo para que se presenten accidentes en el trabajo. Hasta hace algunos años, antes de publicado el presente estudio, el Perú no contaba con una legislatura con rango de ley para regir los estándares de seguridad e higiene industrial básicos y mínimos que prevengan al trabajador de sufrir lesiones o enfermedades producto del trabajo. Hoy la situación es diferente, con la publicación de la Ley de Seguridad y Salud en el Trabajo (Ley 29783) y su reglamento (Decreto Supremo 005-2012-TR) el estado regula,

⁶Saari, Jorma. Accidentes y Gestión de la Seguridad. Enciclopedia de Salud y Seguridad en el Trabajo de la OIT. Pag. 56.3. Disponible en <http://www.insht.es/0057F17F-E555-4734-B4E9-5020B0978F88/FinalDownload/DownloadId-E81887714E078E9BA83EDE7E6D79BEA1/0057F17F-E555-4734-B4E9-5020B0978F88/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

⁷Saari, Jorma. Accidentes y Gestión de la Seguridad. Enciclopedia de Salud y Seguridad en el Trabajo de la OIT. Pag. 56.2. Disponible en <http://www.insht.es/0057F17F-E555-4734-B4E9-5020B0978F88/FinalDownload/DownloadId-E81887714E078E9BA83EDE7E6D79BEA1/0057F17F-E555-4734-B4E9-5020B0978F88/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

fiscaliza y propone medidas basadas crear un sistema de gestión en las empresas que permita controlar los riesgos inherentes a los puestos de trabajo.

1.2.3. Clasificación de los accidentes de trabajo.

Los accidentes de trabajo son clasificados de la siguiente forma según el Reglamento de Seguridad y Salud en el Trabajo⁸:

1. Accidente leve: Suceso cuya lesión, resultado de la evaluación médica, que genera en el accidentado un descanso breve con retorno máximo al día siguiente a sus labores habituales.
2. Accidente incapacitante: Suceso cuya lesión, resultado de la evaluación médica, da lugar a descanso, ausencia justificada al trabajo y tratamiento. Para fines estadísticos no se tomará en cuenta el día de ocurrido el accidente. Según el grado de incapacidad los accidentes de trabajo pueden ser:
 - a. Total temporal: cuando la lesión genera en el accidentado la imposibilidad de utilizar su organismo; se otorgará tratamiento médico hasta su plena recuperación.
 - b. Parcial permanente: cuando la lesión genera la pérdida parcial de un miembro u órgano o de las diferentes funciones del mismo.
 - c. Total permanente: cuando la lesión genera la pérdida anatómica o funcional total de un miembro u órgano; o de las funciones del mismo. Se considera a partir de la pérdida del dedo meñique.
3. Accidente mortal: Suceso cuyas lesiones producen la muerte del trabajador. Para efectos estadísticos debe considerarse la fecha del deceso.

También existe la clasificación informal la cual es como sigue: accidente leve (con menos de 01 día de descanso médico), accidente grave (con más de 01 día de descanso médico) y accidente fatal o mortal (implica muerte del trabajador).

1.2.4. Anuario de Accidentabilidad del Mintra

El Ministerio de Trabajo (MINTRA) publica los anuarios estadísticos⁹ socio laborales a través de su Oficina de Estadística. En estos documentos se registran los índices de accidentabilidad de todo el país. En la publicación correspondiente al año 2010, última publicación según la elaboración del presente trabajo, se presentan los siguientes datos:

1. Verificación de accidentes de trabajo por meses. Según estas cifras se registra un total de accidentes de 277, siendo el mes con la más alta tasa de accidentabilidad, febrero.
2. Verificación de accidentes de trabajo por meses, según direcciones regionales y zonas de trabajo. En este caso la mayor tasa de accidentabilidad la registra la ciudad de Trujillo con 80 accidentes de trabajo.
3. Verificación de accidentes de trabajo por meses, según actividad económica. En esta categoría la industria manufacturera lleva la delantera con 85 casos.

⁸ Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Glosario de términos.

⁹ Ministerio del Trabajo. Anuario Estadístico. Pag. 56.2. Disponible en <http://www.mintra.gob.pe/mostrarContenido.php?id=86&tip=87>

4. Verificación de accidentes de trabajo por mes, según categoría ocupacional. Este dato se refiere al puesto de trabajo. Aquí podemos apreciar que la categoría “otros” tiene la mayor tasa con 131 casos. El operario ocupa el segundo lugar con 67 casos.
5. Verificación de accidentes de trabajo por meses, según forma de accidente. Este punto se refiere a la naturaleza del accidente según el riesgo al cual se encuentra expuesto el trabajador. La categoría “otras formas” es la más alta con una tasa de 55 casos. Por otro lado “aprisionamiento o atrapamiento” ocupa el segundo puesto con 47 casos.
6. Verificación de accidentes de trabajo por meses, según agente causante. Este punto se asocia a la fuente o peligro con potenciales riesgos de generar accidentes de trabajo. La categoría “otros” ocupa el primer lugar con 96 casos seguida de “Máquinas y equipos en general” con 58 casos.
7. Verificación de accidentes de trabajo por meses, según parte de cuerpo lesionada. La categoría “Ubicaciones múltiples, compromiso de dos o más zonas afectadas indicadas en la tabla” ocupa el primer lugar con 55 casos.
8. Verificación de accidentes de trabajo por meses, según la naturaleza de la lesión. Los traumatismos internos son los más frecuentes con 51 casos registrados.
9. Verificación de accidentes de trabajo por meses, según consecuencias del accidente. Los accidentes incapacitantes ocupan el primer lugar con 65% de casos, de los cuales hubieron 102 casos de incapacidad total temporal.

No obstante a estos datos, se sabe que el Perú carece a la fecha, de un sistema integrado de notificación y registro de accidentes laborales de modo que la estadística no refleja necesariamente el comportamiento de la accidentabilidad laboral en el país. Por otro lado, según datos del Banco Mundial a través de la publicación del estudio denominado “El mercado laboral peruano durante el auge y caída”, se determina que el Perú para el año 2008 tenía al menos 11 millones de trabajadores en calidad de informales y 4.1 millones en condición de formalidad¹⁰, esto representa al menos el 70% de la población laboral del Perú. La población laboral informal no goza de los beneficios de ley lo cual, para el caso que ocupa al presente estudio, implica que los accidentes laborales sufridos por trabajadores informales no son adecuadamente registrados por el sistema nacional generando un desconocimiento de las circunstancias, causas y efectos de los mismos por parte del estado.

En este orden de ideas es empíricamente sabido que además en las empresas existe una cultura del sub registro lo que significa que muchos de los accidentes no son registrados o no son reportados a las instancias correspondientes, menos al estado.

Tabla 1.1. **Resumen de estadísticos del anuario 2010 Mintra (accidentes de trabajos por meses)**

Estadístico	Valor
Anual	Febrero
Según direcciones regionales y zonas de trabajo	Trujillo
Según actividad económica	Industria de Manufactura
Según categoría ocupacional	Otros
Según forma de accidente	Otras formas
Según agente causante	Otros
Según parte de cuerpo lesionada	Ubicaciones múltiples, compromiso de dos o más zonas afectadas indicadas en la tabla
Según la naturaleza de la lesión	Traumatismos internos
Según consecuencias del accidente	Accidentes incapacitantes

Fuente: Anuario estadístico 2010 del Mintra

¹⁰ Banco Mundial (2010). *El mercado laboral peruano durante el auge y caída*. Perú: Ledel SAC. Pag. 27.

Se espera que la nueva legislación permita mejorar el proceso de notificación de empresas formales e instituciones públicas, ya que recientemente han sido incluidas gracias a la nueva ley, en el sistema nacional de registro de accidentes laborales. Por otro lado, el tema relativo a la informalidad laboral de los distintos sectores productivos deberá ser tratado con políticas estatales firmes aunque se deduce que la transición continuará siendo prolongada.

1.2.5. La investigación de accidentes

El estudio de la accidentabilidad es un factor clave para entender las causas básicas que los originan. Sobre ellas se establecen mecanismos de control preventivos; es decir, que eviten se concrete el accidente. Para lograr esto es necesario conocer las causas que los originan. La experiencia que se tiene respecto a los accidentes desde los inicios de la revolución industrial hasta la actualidad han permitido conocer patrones comunes de causalidad los cuales podrían agruparse en dos grandes conjuntos: aquellos que están asociados a factores relacionados con el trabajador y aquellos asociados con el empleador.

En ese sentido conviene revisar las teorías más importantes respecto a la causalidad de accidentes.

1. Teoría del dominó, propuesta por Herbert William Heinrich en 1931. Esta teoría divide la causalidad de accidentes en 03 conjuntos: actos humanos peligrosos, condiciones peligrosas y hechos fortuitos en una distribución aproximada de 88%, 10% y 2 % respectivamente. Así mismo, plantea que la secuencia o dinámica de los accidentes se basa en cinco pasos:
 - a. Antecedentes y entorno social.
 - b. Fallo del trabajador.
 - c. Acto inseguro unido a un riesgo mecánico y físico.
 - d. Accidente.
 - e. Daño o lesión

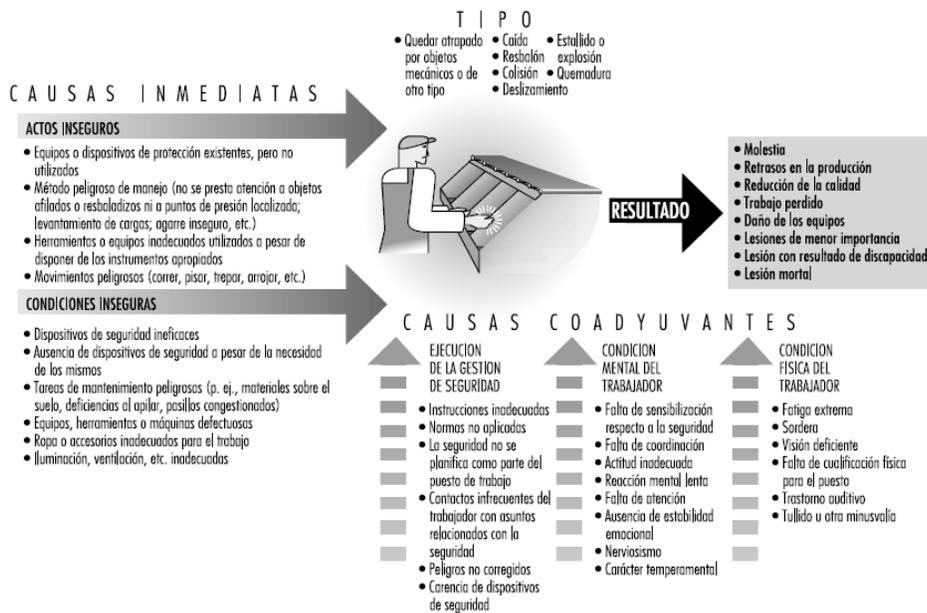
Heinrich planteó que tal como ocurre en una secuencia de fichas de dominó, el hecho de eliminar un paso evitaría la ocurrencia del accidente, de manera similar a cuando luego de retirar una ficha de dominó se evita la secuencia de caída. La teoría del dominó fue pionera en el estudio de causalidad y sentó las bases para las investigaciones futuras.

2. Teoría de la causalidad múltiple, la cual se basa en la teoría del dominó. Básicamente esta teoría sostiene que los accidentes son originados por múltiples causas las cuales se encuentran relacionadas entre sí y que dadas las interacciones entre estas se puede llegar a concretar un accidente. Divide los factores principales de accidentabilidad en dos conjuntos: aquellos asociados al comportamiento del trabajador y aquellos asociados al ambiente de trabajo.
3. Teoría de la causalidad pura, cuyo fundamento se podría considerar apoyado en causalidad estocástica dentro de la cual, la probabilidad de que un trabajador se accidente es la misma y en cuyo caso, las causas obedecen a hechos fortuitos sobre los cuales se destaca la ineficacia de medidas preventivas que los eviten.
4. Teoría de la propensión al accidente, en oposición a la teoría anterior, ésta sostiene que la probabilidad de accidentabilidad en trabajadores no es la misma y que más bien

existen grupos de éstos con probabilidades mayores de concretar accidentes de trabajo. Es una teoría básicamente empírica no basada en datos objetivos.

5. Teoría de transferencia de energía, la cual supone que para que haya un accidente necesariamente tiene que haber transferencia de energía entre un receptor, una fuente y un medio. En todo caso, el objetivo es tratar de controlar estas energías con lo cual se tiene grandes probabilidades de evitar accidentes de trabajo.

Figura 1.2. Estructura de los accidentes



Relación de causas y resultados (accidentes)¹¹

Está claro que la ocurrencia de un evento peligroso el cual eventualmente termine en un accidente, está fuertemente asociada a causas puntuales y posiblemente identificables; sin embargo su deducción es bastante compleja. Por otro lado, la literatura y la experiencia describen muchas causas asociadas al comportamiento del trabajador quién por alguna o muchas razones, no logra identificar ni prevenir factores claves asociados a accidentes laborales.

Al reconocer que los accidentes tienen causas potenciales y sostener esta afirmación como una premisa se tiene como tarea principal desarrollar una serie de técnicas que nos permitan identificar los potenciales factores de accidentabilidad a fin de plantear medidas preventivas. Este proceso altamente técnico el día de hoy, es conocido en términos generales como evaluación y gestión de riesgos. No obstante, a pesar de estos procedimientos y los programas de seguridad industrial que las empresas implementan, los accidentes siguen ocurriendo e inclusive son recurrentes.

La investigación de accidentes es hoy en día una herramienta muy poderosa que permite identificar las causas que originaron el accidente de trabajo; sin embargo se ejecuta cuando

¹¹ Abdul Raouf. Teoría de las causas de accidentes. Pag. 56.6 – 56.7. Disponible en <http://www.insht.es/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

éste ya ocurrió. Aun así es una tarea necesaria y muy influyente en la realización de planes de seguridad industrial. Su desarrollo es básicamente un estudio científico. Sobre este respecto conviene mencionar que existen muchas técnicas para llevar a cabo el proceso. Así mismo y tal como se refirió líneas arriba, existen muchas teorías asociadas a la ocurrencia de accidentes laborales. En ese sentido, las empresas deben desarrollar procedimientos basados en lo siguiente:

1. La legislación.
2. La metodología de investigación.
3. Las teorías de accidentabilidad.

La legislación no provee metodologías específicas pero sí proporciona parámetros importantes a tomar en cuenta en el proceso. Por ejemplo, la legislación peruana suscribe como obligatorio que toda investigación de accidentes deberá ser liderada por el comité de seguridad y salud en el trabajo el cual podría desarrollar comités técnicos de investigación en función a la complejidad del caso.

La metodología de investigación plantea todo el proceso a seguir para hallar las causas fundamentales del accidente. Su correcta definición y ejecución proporcionan grandes probabilidades de encontrar las causas que originaron el accidente y el planteamiento de medidas preventivas integrales. Con respecto a este punto existen muchos estándares internacionales que definen procedimientos muy detallados. El Perú ha publicado el Protocolo para la investigación de accidentes de trabajo¹².

Finalmente, las teorías de accidentabilidad proporcionan los fundamentos o supuestos ya sea con base científica o empírica que explican el por qué se producen los accidentes y que es lo que se tiene que investigar cuando éstos ocurren. Una de las teorías más importantes y de gran aceptación es del Error Humano, el cual estudia las causas por las cuales los trabajadores pueden cometer errores que deriven en sucesos peligrosos. También se tiene el estudio de fallas de equipos y componentes; y finalmente el estudio del medio o entorno de trabajo.

En términos generales y para efectos de investigación de accidentes desarrollados en el Perú, los procedimientos se centran en encontrar causas inmediatas y causas básicas sobre las cuales se plantean medidas de control medibles y objetivas.

1.2.6. Legislación en materia de accidentabilidad laboral

El Perú ha publicado la ley de Seguridad y Salud en el Trabajo 29783 y su reglamento, D.S. 005 2012 TR. Esta legislación suscribe que las empresas e instituciones deberán establecer e implantar un sistema de gestión de seguridad y salud en el trabajo en un marco muy similar al estándar internacional OSHAS 18001. El establecimiento de un sistema de gestión permite a los empleadores tener una visión más integral de la gestión de seguridad y salud en el trabajo con el objeto de evitar accidentes y enfermedades ocupacionales.

En cuanto a la gestión de accidentes laborales, esta legislación ha previsto los siguientes criterios:

1. Mejora Continua.- La investigación de accidentes debe ser básicamente, un *input* para la evaluación y mejora de los procesos que no tengan controles eficientes.

¹²Ministerio de trabajo. *Protocolo para la investigación de accidentes de trabajo*. Disponible en http://www.mintra.gob.pe/archivos/file/dnit/protocolo_investigacion_accidentes_trabajo.pdf

2. Consecuencia de accidentes.- La legislación cubre la posición del trabajador accidentado respecto de posibles indemnizaciones o cambios de puesto de trabajo de ser el caso. Esto es particularmente importante cuando se determina que el trabajador ya no podrá ejercer las funciones que antes tenía.
3. Información de accidentes de trabajo.- Se define la importancia de mantener registros de accidentabilidad y el rol del Ministerio de Trabajo para revisar este tipo de información.
4. Notificación de accidentes de trabajo.- Se suscribe de manera obligatoria que todo accidente de trabajo o incidente deberá ser notificado al Ministerio de Trabajo.
5. Registros de accidentes de trabajo.- Se suscribe la obligatoriedad de contar con registros de accidentabilidad, sus periodos de vigencia y el alcance de los mismos.
6. Investigación de accidentes de trabajo.- Se suscribe que este procedimiento es obligatorio y se imparten pautas generales para realizarlo.

Como se puede apreciar, la legislación es bastante puntual con respecto a la gestión de accidentes, buscando la buena gestión de la información y la eficacia de las medidas adoptadas a fin de evitar la recurrencia de los mismos. Se puede acceder a mayor detalle consultando el anexo B del presente trabajo.

1.3. Minería de Datos

1.3.1. Introducción a la Minería de Datos

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos¹³. La minería de datos trabaja sobre grandes volúmenes de datos que pueden provenir de distintos sistemas de información con el fin de, usando técnicas adecuadas, utilizar los mismos para extraer conocimiento en beneficio de un determinado objetivo.

La minería de datos o *data mining* tiene un amplio espectro de aplicación de donde se pueden citar los siguientes ejemplos: análisis de créditos bancarios, análisis de la cesta de compra, determinar ventas de un producto, determinar grupos diferenciados de empleados, evaluación de campañas publicitarias, descripción y predicción de fallos y accidentes, entre otros.

Los tipos de datos sobre los cuales se puede aplicar minería de datos son muy diversos. En realidad ésta puede aplicarse a cualquier tipo de información sobre las cuales deben usarse distintas técnicas de minería diferentes para cada caso. Los datos pueden estar estructurados como es el caso de las bases de datos relacionales o no, siendo ejemplo de esto último, la información contenida en la *web*.

1.3.2. Business intelligence

Por otro lado, no se puede hablar de minería de datos sin dejar de hablar de *Business Intelligence* o Inteligencia de Negocios traducido al español, el cual compila todas las técnicas que pueden hacer posible la toma de decisiones sobre el análisis de datos relativo a un

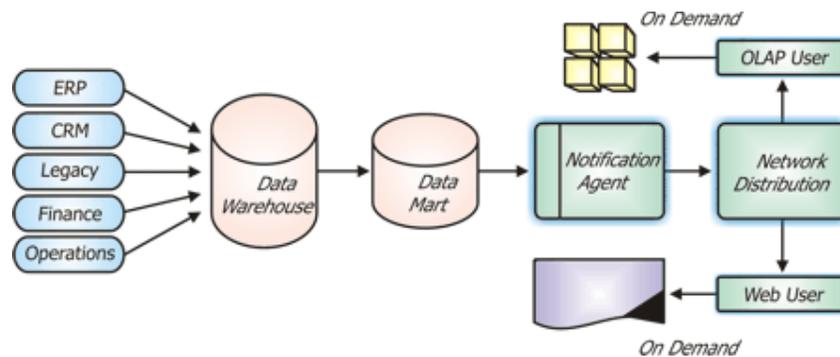
¹³ Hernández José, Ramírez José, Ferri Cesar (2004). *Introducción a la minería de datos*. Madrid: Pearson Prentice Hall. Pag. 05-06.

problema en particular. La minería de datos se presenta como una herramienta que facilita la extracción de información relevante para la toma de decisiones dentro de todo el conjunto de técnicas que conforman un sistema de inteligencia de negocios.

Para el caso de aplicaciones de inteligencia de negocios basados en datos estructurados se tiene que en general los datos están almacenados en un gran repositorio llamado *datawarehouse*, derivado de un sistema de procesamiento *on line* de transacciones (*OLTP*), del cual se puede extraer un conjunto más pequeño de datos, *datamarts*, para posteriormente analizar los datos a través de alguna herramienta de procesamiento analítico *on line* (*OLAP*), un sistema de información para ejecutivos (*EIS*); un sistema de apoyo a la decisión (*DSS*); o aún, un sistema de descubierta y predicción (*data mining*)¹⁴.

A continuación se presenta como ejemplo, un modelo típico de arquitectura de un sistema de inteligencia de negocios basado en datos estructurados.

Figura 1.3. **Arquitectura de un sistema de inteligencia de negocio basado en datos estructurados**



Arquitectura típica de fuentes de datos estructurados¹⁵

1.3.3. Proceso de descubrimiento de conocimiento.

Normalmente la literatura referida a este concepto utiliza se refiere a él como el “descubrimiento de conocimiento en base de datos” (*Knowledge Discovery in Databases KDD*). Siendo sus etapas o fases las siguientes¹⁶:

1. Selección.
 - a. Recopilar e integrar las diversas fuentes de datos existentes.
 - b. Identificar y seleccionar las variables relevantes en los datos.
 - c. Aplicar las técnicas de muestreo adecuadas.
2. Exploración.
 - a. Utilizar las técnicas de análisis exploratorio de datos.
 - b. Deducir la distribución de los datos, simetría y normalidad.
 - c. Analizar las correlaciones existentes en la información.

¹⁴Vieira Luis, Ortiz Luis, Ramírez Santiago (2009). *Introducción a la minería de datos*. Brasil: E-papers Servicios Editoriales. Pag. 20.

¹⁵Fundación Universitaria Iberoamericana FUNIBER. *Business Intelligence y Gestión Documental*. Sección. 1.3.2.2.

¹⁶Pérez César, Santín Daniel (2007). *Minería de Datos Técnicas y Herramientas*. Ed. 01. Madrid: Ediciones Paraninfo S.A. Pag. 06.

3. Limpieza.
 - a. Detectar y tratar la presencia de valores atípicos (outliers).
 - b. Imputar la información faltante o valores perdidos (datos missing).
 - c. Eliminar datos erróneos o irrelevantes.
4. Transformación.
 - a. Utilizar técnicas de reducción y aumento de la dimensión.
 - b. Aplicar técnicas de discretización y numeración.
 - c. Realizar escalado simple y multidimensional.
5. Minería de datos.
 - a. Utilizar técnicas predictivas.
 - b. Utilizar técnicas descriptivas.
6. Evaluación e interpretación de resultados.
7. Difusión y uso de modelos.

1.3.4. Técnicas de Minería de Datos.

Las técnicas de minería de datos se usan en función de la tarea o problema a resolver, como lo es encontrar patrones de accidentabilidad. Este caso en particular representa una tarea de agrupamiento o *clustering* que puede ser resuelta usando redes neuronales, el algoritmo *Kmeans*, entre otras.

A saber, las principales tareas de minería de datos son: Predictivas y descriptivas; mientras que en función de estas tareas se tienen las siguientes técnicas de minería de datos, como más comunes e importantes: Técnicas algebraicas y estadísticas, técnicas bayesianas, técnicas basadas en conteos de frecuencias y tablas de contingencia, técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas, técnicas relacionales, descriptivas y estructurales, técnicas basadas en redes neuronales artificiales, técnicas basadas en núcleo y máquinas de soporte vectorial, técnicas estocásticas y difusas; y finalmente, técnicas basadas en casos, en densidad o distancia.

La clasificación anterior es referente pues la literatura muestra tipos más o menos parecidos, además de híbridos lo que dificulta la taxonomía.

Por otro lado, las principales técnicas de minería de datos se suelen clasificar según su tarea de descubrimiento en¹⁷: agrupación o *clustering*, clasificación y asociación.

La agrupación o *clustering* corresponde a una tarea descriptiva y su objetivo es obtener grupos o conjuntos de los cuales sus elementos tienen características en común. En la agrupación se desconoce *a priori* las características de los grupos. Uno de los algoritmos de *clustering* más usado es *K-means*.

La clasificación corresponde a una tarea predictiva que permite clasificar un conjunto de datos de acuerdo a una premisa o etiqueta. Un ejemplo de esto se ve en los supermercados en donde dado un conjunto de características del comprador se le clasifica en “compra el producto” y “no compra el producto”. A diferencia del *clustering*, en este caso sí se tiene pleno conocimiento de los conjuntos o clases que se pretenden analizar y de la regla o etiqueta que

¹⁷Perversi, Ignacio (2007). *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina*. Tesis de grado en ingeniería industrial. Buenos Aires: Instituto Tecnológico de Buenos Aires. Pag. 04.

permitirá realizar la clasificación de elementos. Los modelos de clasificación se construyen en función de un clasificador sobre un conjunto de entrenamiento el cual dado un método supervisado (se conoce la clase de pertenencia y se le indica al modelo si la clasificación hecha es correcta o no) analiza los datos de entrenamiento a fin de elaborar una descripción para cada clase utilizando las características disponibles de los datos. De este modo se podrá clasificar otras instancias para las cuales se desconoce su clase. Se utilizan algoritmos de inducción para este tipo de tareas.

La asociación corresponde a una tarea descriptiva que permite generar relaciones entre los valores de los atributos de dos o más conjuntos de datos. Por ejemplo para el caso del comprador de supermercado se podría tener la siguiente regla de asociación: “Si edad comprador = 40 años \wedge sueldo trabajador \geq 2500 soles entonces compra bicicleta = Sí”. El algoritmo *A priori* es el más usado para este tipo de tarea.

A continuación se presenta una tabla conteniendo un resumen de las técnicas de minería de datos según la tarea propuesta:

Tabla 1.2. Técnicas de minería de datos

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones/ Factorizaciones
Redes neuronales	X	X	X		
Árboles de decisión ID3, C4.5, C5.0	X				
Árboles de decisión CART	X	X			
Redes Kohonen			X		
Regresión lineal y logarítmica		X			X
Regresión logística	X			X	
<i>Kmeans</i>			X		
<i>A priori</i>				X	
<i>Naive Bayes</i>	X				
Vecinos más próximos	X	X	X		
Algoritmos genéticos y evolutivos	X	X	X	X	X

Fuente: Adaptación del autor¹⁸.

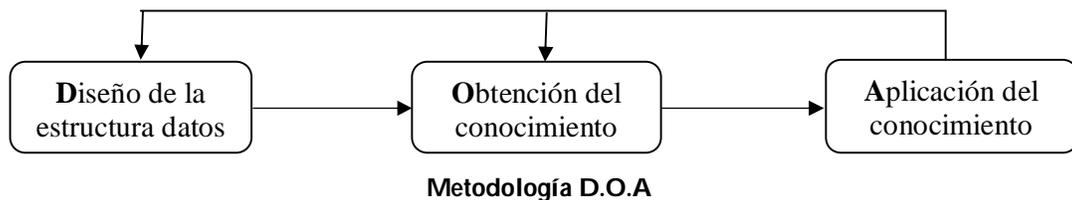
¹⁸ Hernández José, Ramírez José, Ferri Cesar (2004). *Introducción a la minería de datos*. Madrid: Pearson Prentice Hall. Pag. 148.

Capítulo 2

Guía metodológica para obtener patrones de accidentabilidad laboral basado en *data mining*

A continuación se expone de manera detalla, la metodología a seguir para analizar datos de accidentabilidad laboral y obtener patrones o conjuntos de datos comunes que puedan aportar conocimiento conciso del comportamiento o dinámica de los mismos datos. A continuación se sugiere un esquema genérico de la metodología para obtener patrones de accidentabilidad partiendo del diseño de la estructura de datos, ejecución del *data mining* para obtener conocimiento y la aplicación del mismo en los programas preventivos en seguridad industrial. Para efectos nemotécnicos se denominará como D.O.A a todo el esquema metodológico, la cual en términos prácticos es una versión resumida del *KDD*.

Figura 2.1. Esquema de la metodología D.O.A



El proceso inicia con el **Diseño de una estructura de datos** adecuada, esto es que sea correcta desde el punto de vista técnico y legal. Este último punto es muy importante pues la información que se vaya a analizar no tiene por qué generar discrepancias o dificultades con respecto a la información que la legislación peruana exige.

La base de datos surge básicamente de lo que establece la legislación a través de los registros de accidentabilidad obligatorios y de la experiencia que tenga cada empresa. Aquí es donde los procedimientos de investigación de accidentes cobran mucha relevancia, ya que generan *inputs* de datos bajo ciertos supuestos o situaciones, que podrían analizarse y enriquecer no sólo los resultados de la investigación sino la misma base de datos, con lo cual su crecimiento no solo sería vertical sino también horizontal (se pueden generar nuevos campos). Sin embargo, es sabido que lo más importante para una investigación de accidentes es hallar las causas básicas relacionadas con los accidentes a fin de establecer medidas de mejora continua que sean integrales de modo que el análisis de los datos también se comporta como *input* para

la investigación lo que convierte el proceso general en un beneficioso sistema interdependiente de información el cual deberá generar conocimiento para la empresa.

Con una base de datos adecuada se puede realizar la **obtención de conocimiento**, lo que técnicamente sería el siguiente gran paso a realizar, mediante las tareas de minería de datos necesarias para hallar patrones de accidentabilidad. Ahora bien, la base de datos es una herramienta de almacenamiento dinámico y en este caso es crucial que así sea ya que el análisis de los datos debe realizarse desde diferentes puntos de vista, contextos o escenarios. En este punto, la extracción de *datasets* o tablas específicas derivadas de la base de datos madre, resulta muy oportuno, ya que con ella se realizarán las tareas de *data mining* que el analista requiera.

Realizado el análisis con las técnicas de minería de datos apropiadas de acuerdo a lo que se pretenda encontrar, se deberá someter a juicio de expertos los resultados obtenidos a fin de compararlo con la realidad. Este paso es vital a fin de evitar falsos supuestos y sobre todo, someter todo el proceso a mejora continua con lo cual se pueden deducir nuevas líneas de investigación, nuevas formas de ver el comportamiento de los datos, identificar procesos que no tiene control y que se relacionan con accidentes, etc. En general, obtener más y más conocimiento.

El estudio de patrones de accidentabilidad es tratar de hallar regularidades en los datos que tengan relación directa con la ocurrencia de accidentes; esto es, con sus causas básicas. Por eso es de vital importancia identificar los campos que permiten esta relación, ya que en adelante se convertirán en las variables de estudio de los patrones de accidentabilidad.

Finalmente se tiene la **aplicación del conocimiento** lo cual ciertamente es lo más importante de todo el proceso. El conocimiento obtenido debe permitir tomar decisiones para mejorar los programas preventivos en materia de seguridad y salud en el trabajo así como mejorar la experiencia en las técnicas de investigación de accidentes, desde el trabajo de análisis hasta el registro en la base de datos. Son estos dos campos los más importantes a considerar al momento de hallar utilidad clara del conocimiento resultante de los patrones de accidentabilidad laboral. Por ejemplo, es sabido que muchos accidentes asociados a riesgo eléctrico están relacionados a problemas con los sistemas de bloqueo y etiquetado de las fuentes generadoras de electricidad o de los paneles de control. Eventualmente un tercero no relacionado con el trabajo, procede a habilitar la llave que se desactivó para realizar el trabajo, generando la realimentación del circuito con la frecuente consecuencia de generar un shock eléctrico al trabajador que realiza su labor en un punto dado del mismo. En este caso el sistema de bloqueo falló y el investigador intentará descubrir el por qué. Ahora bien, es importante notar que no basta con decir “el sistema falló” pues es necesario encontrar que otras variables se asocian a este problema y que en conjunto, generan la cadena de eventos que termina en el accidente de shock eléctrico. En las conclusiones de la investigación quizá se encuentren variables que al parecer no podrían tener mucha relación tales como el día, la hora, el mes, el número de recurrencias del accidente, la competencia del trabajador, etc. Es aquí donde el estudio de los datos puede revelar información escondida y de hecho muy relevante para un sistema de prevención de seguridad laboral.

2.1. Diseño de una estructura de datos de accidentabilidad laboral

Para construir una base de datos de accidentabilidad laboral, se puede usar cualquier herramienta que se encuentre al alcance. Al respecto se pueden hacer varias recomendaciones

sobre el tipo de tecnología a utilizar; sin embargo ello no es el propósito del presente estudio. En todo caso es necesario decidir el modelo de base de datos y el sistema de gestión de bases de datos a usar. Lo importante es modelar la base de datos en función de lo requerido por la legislación peruana y la información que se quiera registrar de manera complementaria.

Con respecto a la información complementaria conviene realizar algunos comentarios:

1. Dicha información resulta principalmente de los procesos de mejora continua obtenido en las investigaciones de accidentes laborales.
2. Esta información puede hallarse como deducción o combinación de una serie de supuestos o escenarios que el analista estudia a fin de comprobar ciertas conjeturas o en su defecto, encontrar características pre asumidas.
3. Los motores de gestión de bases de datos ayudan actualmente mucho a esta labor.
4. La legislación peruana no especifica ninguna obligación respecto al modo de gestión de los datos de tal forma que el diseño de la misma queda a voluntad del analista. Ciertamente se tiene que cumplir como mínimo con registrar los datos requeridos por la norma.

Para iniciar el estudio de patrones de accidentabilidad, el analista deberá construir un *dataset* que contenga los campos o mejor aún, las variables, que se quieran estudiar. Este *dataset* también puede ser llamado como vista minable y resulta de un proceso de selección, exploración y transformación de datos a partir de una base de datos de accidentabilidad previamente definida.

2.1.1. Procedimiento de investigación de accidentes

Se había visto en la sección 1.2.5 las generalidades más importantes a considerar en un protocolo de investigación de accidentes. Actualmente, existen muchas técnicas y metodologías para hacerlo, inclusive se utilizan técnicas propias de la criminología cuando se quieren estudiar fatalidades. El analista puede escoger las herramientas, tal cual ocurre para el caso de las bases de datos, que considere oportunas sin embargo deberá reconocer que cada industria tiene sus propias características lo cual hace que las variables de estudio puedan ser diferentes. Dada esta razón conviene revisar las cuestiones más relevantes del proceso a fin de aprovechar al máximo la extracción de variables de estudio.

1. La legislación peruana suscribe que es necesario conformar un comité de seguridad integrado por representantes del empleador y del trabajador junto a organizaciones sindicales, si en caso las hubieren, para realizar la investigación¹⁹.
2. En caso se requiera asistencia técnica la legislación propone la posibilidad de conformar comités técnicos²⁰. Esto es particularmente importante, dado que se pueden obtener datos muy específicos respecto a situaciones concretas. Por ejemplo, si una batería de un generador eléctrico explota al parecer bajo condiciones normales, conviene saber qué factores de su estructura interna fallaron. Los peritajes son los métodos más solicitados en estos casos.
3. Los datos recogidos a través de notas, fotografías, audios y/o videos deben ser analizados y registrados considerando los datos que deben registrarse en la base de

¹⁹ Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Artículo 42, literal 1'.

²⁰ Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Artículo 65.

datos y los que pueden soportarse en otro medio, como lo es el informe de conclusiones. Recuérdese que la idea no es atosigar la base de datos haciéndola crecer desmedidamente de forma horizontal ya que cada caso es diferente y las bases de datos almacenan generalidades, a menos que se diseñen modelos particulares a gusto del analista. Por ejemplo, no se requerirá conocer el amperaje al cual fue expuesto la batería que explotó en el caso anterior como campo o dato en otra investigación de accidentes.

4. El primer gran paso del *output* de la investigación es hallar las causas fundamentales, directas o básicas (como quiera llamársele) asociadas al accidente de trabajo, ya que estas pueden representar tendencias o patrones comunes registrados en diversos accidentes. Por ejemplo, es sabido que muchos accidentes ocurren debido a que los trabajadores no usan sus equipos de protección personal; sin embargo, podría ser que los casos más frecuentes ocurran cuando se va a terminar la jornada, de lo cual se puede deducir a priori que el trabajador podría dejar de usarlos precisamente en esas horas. Como se ve, alrededor de la causa básica (no usar el equipo de protección personal) hay un variable que podría resultar interesante (hora en que los trabajadores deciden retirarse sus equipos de protección personal).
5. Las variables asociadas a las causas básicas no son fácilmente asociables, están allí pero no dicen que tiene que ver con tal o cual causa. Justamente por eso se recurre al *data mining*. Es importante considerar con criterio y apoyados en la opinión del personal experto, protagonista y experimentado los detalles que podrían ser relevantes. Por ejemplo, es fácil asociar al cansancio los accidentes laborales que ocurren poco antes de terminar la jornada laboral pero podría ser poco evidente que esto es más frecuente en verano que en invierno. Podría resultar que la estación del año marque un patrón de frecuencia.
6. El segundo gran *output* de la investigación son las lecciones aprendidas puesto que se convierten en premisas de acción para los planes de seguridad industrial pero más aún, generan escenarios probables para analizar supuestos con las variables. Por ejemplo para el caso del accidente asociado a riesgo eléctrico, se concluyó como causa básica, la falta de bloqueo de las tableros de control eléctrico, la lección aprendida es que en adelante se deberán bloquear de todas formas pero lo que en realidad está diciendo el caso es que hay una falta de control por parte del responsable del trabajo lo cual podría ocurrir en otro escenario, es un mensaje de alerta. La falta de control es información deducible a partir de varios datos, es una conclusión.

Lo importante de este procedimiento para el estudio de datos de accidentabilidad y su dinámica es que precisamente la investigación de accidentes se comporta como *input* muy valioso del estudio. Proporciona datos sobre las causas básicas que originaron el accidente laboral y es eso lo que no se puede dejar escapar. Esto ya se había tocado anteriormente, también se había referido que la importancia suprema del proceso es encontrar variables asociadas directamente a las causas de los accidentes a tal punto que incluso sean en sí mismas causas directas de accidentes.

2.1.2. Gestión de la data de accidentabilidad

El primer gran bloque del proceso es tener lista una base de datos de accidentabilidad. Se han considerado dos criterios muy importantes: la legislación y el contexto. Entiéndase como contexto las situaciones particulares de la industria que merecen la pena ser registrados a juicio del analista o analistas respecto a las características de los accidentes registrados.

2.1.2.1. Modelo de datos según lo requerido por la legislación peruana.

Como se ha mencionado anteriormente, la legislación peruana ha publicado la ley 29783 de Seguridad y Salud en el trabajo y su respectivo reglamento Decreto Supremo D.S. 005-2012-TR, siendo éste último el que proporciona mayor información sobre qué datos registrar.

Adicionalmente, el presente año el Ministerio de Trabajo y Promoción del Empleo ha publicado la Resolución Ministerial R.M. 050-2013-TR, la cual define a través de su artículo 1, los formatos con la información mínima que las empresas e instituciones deben registrar respecto a sus sistemas de gestión de seguridad y salud en el trabajo. Dentro de este marco, se encuentran los formatos propuestos para registrar accidentes de trabajo. El artículo 2 de esta normativa suscribe que dichos formatos son referenciales; sin embargo, debe quedar claro que los campos planteados son los que como mínimo se deberían registrar.

La ley 29783 menciona en su artículo 87 que los empleadores deberán contar con registros de accidentes de trabajo. El D.S. 005-2012-TR suscribe en su artículo 33 los registros obligatorios del sistema de gestión de seguridad y salud en el trabajo. El literal 'a' de dicho artículo suscribe el registro relacionado a accidentes de trabajo. La R.M. 050-2013-TR contiene el modelo de registro referido en dicho literal. Este registro se mantiene a fin de contener información relevante del accidente de trabajo según el proceso de investigación de accidentes que se haya realizado.

Por otro lado, el D.S 005-2012-TR dispone a través de su artículo 111 que la notificación de accidentes y enfermedades ocupacionales es obligatoria a través del portal web del Ministerio de Trabajo²¹. Para acceder, los empleadores deberán tener activos sus usuarios y claves SOL. Para casos de zonas que no dispongan de servicio de internet, la notificación se deberá hacer a través de documentos en físico según los formatos obligatorios de notificación de accidentes de trabajo y enfermedades ocupacionales según los anexos 01 y 02 de éste reglamento.

Figura 2.2. Registro y notificación de accidentes de trabajo

²¹Ministerio de trabajo. *Sistema de accidentes de trabajo*. Disponible en <https://www.sunat.gob.pe/xssecurity/SignOnVerification.htm?signonForwardAction=http://luna.mintra.gob.pe/si.sat/index.jsp>



Principales categorías de datos en el registro y notificación de accidentes de trabajo

Se pueden apreciar dos procesos claros con respecto a la gestión de datos de accidentabilidad: por un lado se debe tener registro de la investigación del accidente con datos detallados sobre el mismo y por otro, se debe tener registro de la notificación del accidente, con información resumida y objetiva de acuerdo a lo dispuesto por el D.S. 005-2012-TR en sus artículos 111 y 112 respectivamente. Lo que debe ser de interés del analista, es el primer proceso, ya que sobre ese se va a obtener información relevante o conocimiento dentro del proceso *KDD*.

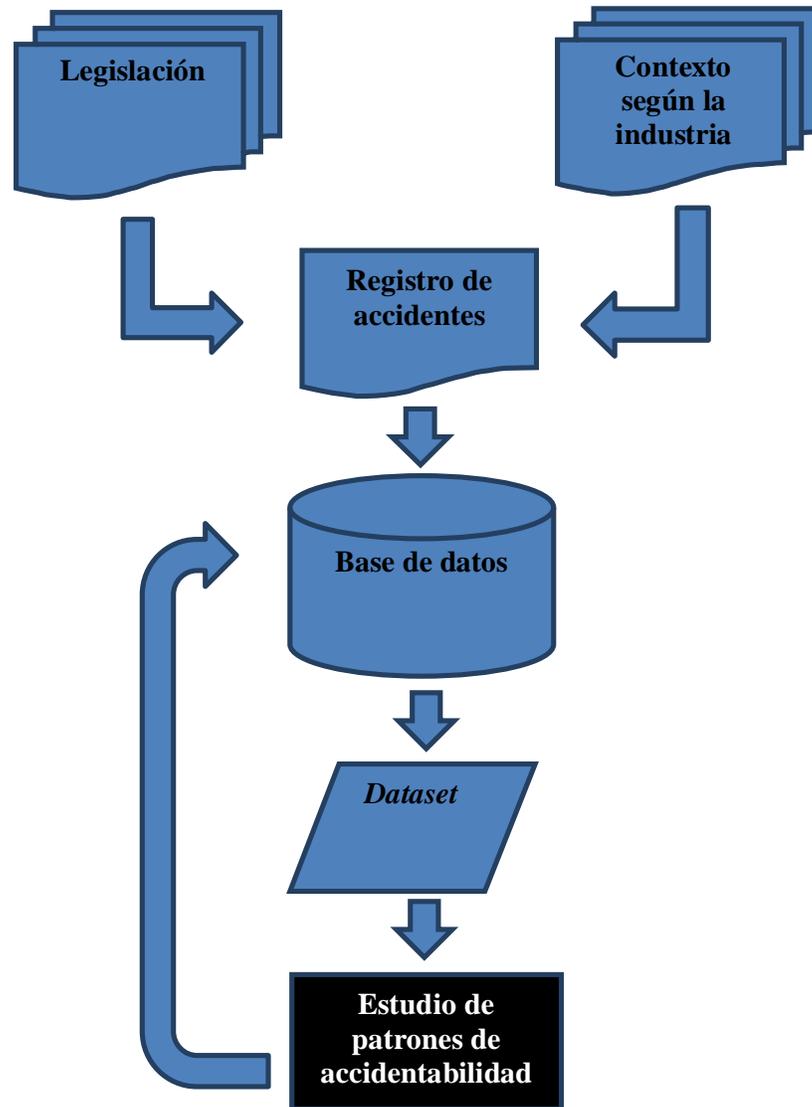
Es preciso aclarar que lo primero que el empleador debe hacer ante la ocurrencia de un accidente de trabajo es la notificación al Ministerio de Trabajo a través de los medios antes comentados. Resulta obvio que para realizar la notificación, el empleador deberá tener al menos datos mínimos según lo establece el D.S. 005-2012-TR; sin embargo no se conocen aún, las causas y variables que originaron el accidente laboral. Cuando el proceso de investigación de accidentes haya terminado, se utiliza el registro de accidente de trabajo y se actualiza la base de datos según los campos definidos. El Ministerio de Trabajo podrá realizar las inspecciones y auditorías a estos registros según sus protocolos de inspección.

2.1.2.2. Datos relevantes.

Tanto para diseñar la base de datos de accidentes de trabajo como para realizar el estudio de patrones es necesario conocer que datos irán en una y otra estructura (base de datos y *dataset* o vista minable). Ya se ha comentado, que la idea es derivar un *dataset* de la base de datos para que sirva de *input* al proceso de estudio de patrones de accidentabilidad. A saber, este *dataset* es una versión resumida de la base de datos con los campos considerados como relevantes para el estudio, esto es: aquellos que posiblemente tienen relación directa con la ocurrencia de un accidente de trabajo.

Ahora bien, conviene plantear una vista transversal del proceso a este nivel. Se pueden apreciar dos procesos claros con respecto a la gestión de datos de accidentabilidad: por un lado se debe tener registro de la investigación del accidente con datos detallados sobre el mismo y por otro, se debe tener registro de la notificación del accidente, con información resumida y objetiva de acuerdo a lo dispuesto por el D.S. 005-2012-TR en sus artículos 111 y 112 respectivamente. Lo que debe ser de interés del analista, es el primer proceso, ya que sobre ese se va a obtener información relevante o conocimiento dentro del proceso *KDD*.

Figura 2.3. Estructura de datos de accidentabilidad



Flujo de comunicación de la base de datos y el estudio de patrones

En el gráfico anterior se puede observar cómo se constituye la base de datos y cómo es su relación con el estudio de patrones de accidentabilidad, asumiendo que este proceso es una caja negra, cuyos detalles se verán más adelante.

Es interesante observar como los resultados alcanzados en el proceso de obtención de patrones de accidentabilidad podrían permitir mejoras en la base de datos haciendo que ésta incorpore nuevos campos en su estructura. En esa línea de ideas, también se puede esperar que las mejoras obtenidas en el estudio de patrones permita mejorar los registros de accidentabilidad sin alterar lo mínimo pedido por la ley por supuesto.

En relación a lo anterior conviene reforzar la idea que para incorporar campos en la base de datos de accidentabilidad estos deberán tener relación con todos los casos que pudieran presentarse, caso contrario podría generar problemas a lo interno de la base de datos. Por ejemplo, la hora en que ocurre un accidente es un hecho que se presenta en cualquier caso, pero la marca de una batería que generó un accidente, solo importa en ese caso en particular, a menos que se analicen casos homogéneos o que el analista encuentre una forma de incorporar cualquier tipo de datos sin afectar la eficiencia de la base de datos.

A continuación se analizará el primer gran bloque de datos a considerar, aquellos que pide mínimamente la ley. Recuérdese que el objetivo en este punto, es modelar la base de datos.

Tomándose como referencia el anexo 1 de la R.M. 050-2013-TR se tiene:

1. Datos del empleador principal.- Este grupo de datos son de naturaleza general y no serán necesarios para el estudio de patrones de accidentabilidad ya que no tienen relación directa u objetiva con la ocurrencia de accidentes.
2. Datos del empleador de intermediación, tercerización, contratista, subcontratista, otros.- *Ídem* que el caso anterior.

Figura 2.4. Datos referidos al empleador

DATOS DEL EMPLEADOR PRINCIPAL:																			
1	RAZÓN SOCIAL O DENOMINACIÓN SOCIAL	2	RUC	3	DOMICILIO (Dirección, distrito, departamento, provincia)	4	TIPO DE ACTIVIDAD ECONÓMICA	5	Nº TRABAJADORES EN EL CENTRO LABORAL										
6					COMPLETAR SÓLO EN CASO QUE LAS ACTIVIDADES DEL EMPLEADOR SEAN CONSIDERADAS DE ALTO RIESGO														
Nº TRABAJADORES AFILIADOS AL SCTR		Nº TRABAJADORES NO AFILIADOS AL SCTR		NOMBRE DE LA ASEGURADORA															
Completar sólo si contrata servicios de intermediación o tercerización:																			
DATOS DEL EMPLEADOR DE INTERMEDIACIÓN, TERCERIZACIÓN, CONTRATISTA, SUBCONTRATISTA, OTROS:}																			
7	RAZÓN SOCIAL O DENOMINACIÓN SOCIAL	8	RUC	9	DOMICILIO (Dirección, distrito, departamento, provincia)	10	TIPO DE ACTIVIDAD ECONÓMICA	11	Nº TRABAJADORES EN EL CENTRO LABORAL										
12										COMPLETAR SÓLO EN CASO QUE LAS ACTIVIDADES DEL EMPLEADOR SEAN CONSIDERADAS DE ALTO RIESGO									
Nº TRABAJADORES AFILIADOS AL SCTR			Nº TRABAJADORES NO AFILIADOS AL SCTR			NOMBRE DE LA ASEGURADORA													

Campos generales que resultan poco relevantes

Resulta claro que agregar estos campos en la base de datos no va a resultar muy conveniente ya que no se van a generar registros diferentes por cada accidente presentado a menos que se traten situaciones especiales que el analista deberá revisar. Consecuentemente los campos no van a ser necesarios para el *dataset*.

3. Datos del trabajador.- Estos campos sí deberían registrarse en la base de datos y por tanto deben formar parte del modelo de datos que el analista proponga.

Figura 2.5. Datos referidos al trabajador

DATOS DEL TRABAJADOR:								
13 APELLIDOS Y NOMBRES DEL TRABAJADOR ACCIDENTADO:					14 N° DN/ICE		15 EDAD	
16	17	18	19	20	21	22	23	
ÁREA	PUESTO DE TRABAJO	ANTIGÜEDAD EN EL EMPLEO	SEXO F/M	TURNO D/T/N	TIPO DE CONTRATO	TIEMPO DE EXPERIENCIA EN EL PUESTO DE TRABAJO	N° HORAS TRABAJADAS EN LA JORNADA LABORAL (Antes del accidente)	

Campos importantes de incluir en el modelado de la base de datos

Es importante notar que hay varios campos que podrían incluirse en el dataset para estudio de patrones de accidentabilidad tales como: edad, área, puesto de trabajo, antigüedad en el empleo, sexo, turno, experiencia y horas trabajadas en la jornada laboral antes del accidente.

Los nombres y el tipo de contrato resultan a priori sin relación directa con ocurrencia la de accidentes, por tanto se descartarían para el dataset aunque sí deberían estar en la base de datos.

4. Investigación del accidente de trabajo.- Estos campos sí deberían registrarse en la base de datos, aunque no todos podrían considerarse en el dataset para estudio de patrones de accidentabilidad.

La fecha y hora del accidente junto al lugar donde éste ocurrió sí podrían considerarse en la base de datos y en el dataset ya que podrían tener relación directa con la ocurrencia del accidente, recuérdese que estos datos se relacionan a situaciones previas al mismo, mientras que los otros campos se registran luego de ocurrido el accidente.

Figura 2.6. Datos referidos a la investigación del accidente de trabajo

INVESTIGACIÓN DEL ACCIDENTE DE TRABAJO										
24 FECHA Y HORA DE OCURRENCIA DEL ACCIDENTE				25 FECHA DE INICIO DE LA INVESTIGACIÓN			26 LUGAR EXACTO DONDE OCURRIÓ EL ACCIDENTE			
DÍA	MES	AÑO	HORA	DÍA	MES	AÑO				
27 MARCAR CON (X) GRAVEDAD DEL ACCIDENTE DE TRABAJO				28 MARCAR CON (X) GRADO DEL ACCIDENTE INCAPACITANTE (DE SER EL CASO)				29 N° DÍAS DE DESCANSO MÉDICO		30 N° DE TRABAJADORES AFECTADOS
ACCIDENTE LEVE	ACCIDENTE INCAPACITANTE	MORTAL	TOTAL TEMPORAL	PARCIAL TEMPORAL	PARCIAL PERMANENTE	TOTAL PERMANENTE				
31 DESCRIBIR PARTE DEL CUERPO LESIONADO (De ser el caso):										

Campos importantes de incluir en el modelado de la base de datos

Con respecto al lugar en donde ocurre el accidente, se deberá tener especial cuidado ya que mucho nivel de detalle puede traer problemas al momento de configurar el modelo de datos e incluso cuando se defina el dataset. Por ejemplo un accidente podría ocurrir en la zona de cargas del taller de mecánica I de la planta central de la estación III de la empresa. El analista deberá evaluar el grado de disgregación en la base de datos y dataset. Deberá evaluar si conviene estudiar los accidentes por estaciones de trabajo o por espacios de trabajo de sus diversos talleres.

5. Descripción del accidente de trabajo.- Se refiere a la descripción objetiva de los hechos o circunstancias en las que ocurre el accidente. Este campo es importante; sin embargo suele estar cargado de mucho texto y eso podría ser un problema para la base de datos. Por otro lado la información que allí se contiene se puede normalizar para utilizar los datos que se describen y registrarlos en campos apropiados. Por ejemplo, la descripción contendrá el lugar, la hora, la fecha, el turno de trabajo, etc. Estos datos pueden ser registrados en sus respectivos campos. El analista deberá de considerarlo en el modelo de datos.

Figura 2.7. Datos referidos a la descripción de hechos del accidente de trabajo

32 DESCRIPCIÓN DEL ACCIDENTE DE TRABAJO
<p>Describe sólo los hechos, no escriba información subjetiva que no pueda ser comprobada.</p> <p>Adjuntar:</p> <ul style="list-style-type: none"> - Declaración del afectado sobre el accidente de trabajo. - Declaración de testigos (de ser el caso). - Procedimientos, planos, registros, entre otros que ayuden a la investigación de ser el caso.

Campos importantes de incluir en el modelado de la base de datos

6. Descripción de las causas que originaron el accidente de trabajo.- Según refiere esta parte de la resolución en cuestión (documento legal), cada empleador tiene la libertad de utilizar el método que mejor le convenga para determinar y describir las causas que originaron el accidente. Ya se había visto en el capítulo 1 del presente trabajo, que existen muchas formas de clasificar esta información pero que en general, el resultado siempre será la lista de causas fundamentales que deben controlarse o corregirse para evitar la recurrencia del accidente. En ese sentido conviene hacer algunas acotaciones puntuales con respecto al estudio de patrones de accidentabilidad. Las causas fundamentales de un accidente son hechos que en resumidas cuentas ocurren antes que éste suceda. En el sentido estricto, algo que debería haberse hecho de un modo se obvia o se hace de otro, habiendo una razón o causa básica que motiva a ese hecho, generando el accidente de trabajo. El analista tiene toda la posibilidad de hallar patrones de repetición de causas pero tiene que tener cuidado en realizar una correcta clasificación pues la interpretación de los resultados puede inducir al error. En algunos casos podría resultar mejor deducir las causas a partir del estudio de patrones de variables independientes. En todo caso, estas decisiones se toman cuando se realiza el estudio y dependiendo del objeto de estudio. Esta información sí debe registrarse en la base de datos aunque no necesariamente puede considerarse en el dataset.

Figura 2.8. Datos referidos a las causas del accidente de trabajo

33 DESCRIPCIÓN DE LAS CAUSAS QUE ORIGINARON EL ACCIDENTE DE TRABAJO
<p>Cada empresa o entidad pública o privada, puede adoptar el modelo de determinación de causas, que mejor se adapte a sus características y debe adjuntar al presente formato el desarrollo de la misma.</p>

Campos importantes de incluir en el modelado de la base de datos

7. Medidas correctivas.- Las medidas correctivas sí deben registrarse en la base de datos pues la medida de su eficacia es lo que será evaluado por el Ministerio de Trabajo. Por otro lado, la información que aporta el estudio de causas básicas y el estudio de patrones de accidentabilidad permitirá decidir las mejores medidas correctivas. Esta información no debería considerarse en el dataset ya que no representa información a considerar antes del accidente.

Figura 2.9. **Medidas correctivas del accidente de trabajo**

34 MEDIDAS CORRECTIVAS					
DESCRIPCIÓN DE LA MEDIDA CORRECTIVA	RESPONSABLE	FECHA DE EJECUCIÓN			Completar en la fecha de ejecución propuesta, el ESTADO de la implementación de la medida correctiva (realizada, pendiente, en ejecución)
		DÍA	MES	AÑO	
1.-					
2.-					

Insertar tantos renglones como sean necesarios.

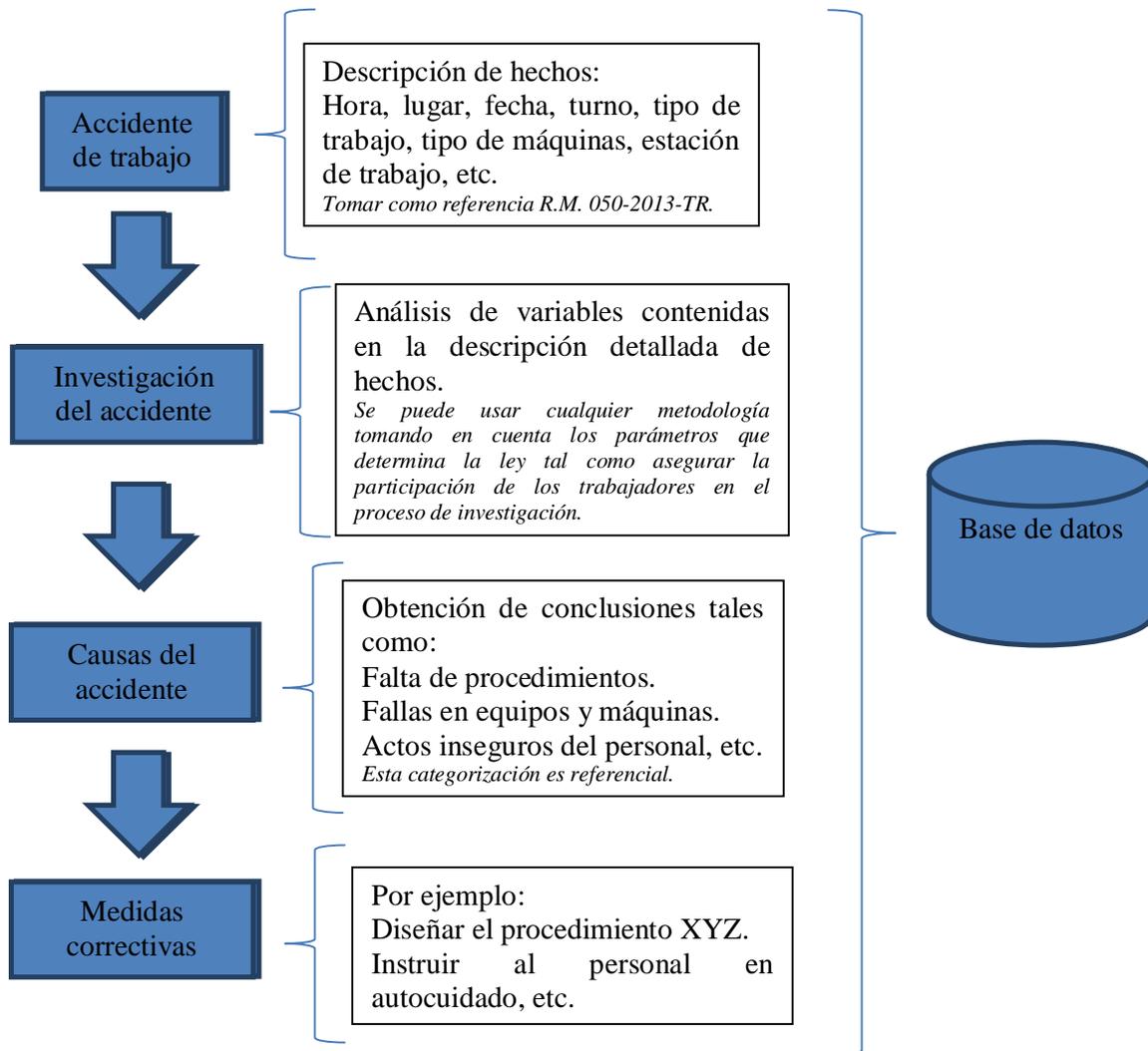
Campos importantes de incluir en el modelado de la base de datos

El documento antes revisado termina con el registro de los responsables de la investigación, lo cual sería prudente registrar en la base de datos.

Ciertamente cualquier dato es susceptible de ser registrado en la base de datos; sin embargo no todos ellos serán importantes de considerar en el *dataset*. Así mismo hay considerar que cada accidente de trabajo generará un registro dentro de la base de datos. Esto es importante si queremos optimizar el rendimiento de la misma.

En este punto se puede realizar un esquema de ideas para visualizar el proceso de generación de datos susceptibles de estudio:

Figura 2.10. **Proceso de investigación de accidentes de trabajo**



Datos obtenidos en la investigación que pueden registrarse en la base de datos

En el modelo gráfico anterior se puede apreciar el proceso de investigación de accidentes y la posibilidad de poder registrar todos los datos que de allí se obtengan en la base de datos.

2.1.2.3. Análisis descriptivo de datos en relación al modelo de gestión de la empresa.

El análisis descriptivo de datos de accidentabilidad puede realizarse con los métodos clásicos estadísticos o con uso del *data mining*. A saber, este último es mucho más potente que los métodos clásicos pues la idea es obtener conocimiento objetivo para la toma de decisiones. En el caso de la accidentabilidad, la toma de decisiones está asociada a las medidas correctivas las que se suponen deben ser eficaces para evitar la recurrencia de accidentes de trabajo.

Para saber que datos son los más destacados de evaluar a través de métodos clásicos, tales como el estudio de frecuencias o modas, bastaría con consultar bibliografía de seguridad industrial o los reportes de estadística de accidentabilidad laboral de instituciones

internacionales como la Organización Internacional del Trabajo o el anuario nacional del Ministerio de Trabajo, cuyo resumen 2010 se encuentra descrito en el capítulo 1.2.4 del presente trabajo. En general, los resúmenes estadísticos se centran en conocer datos tales como: qué parte del cuerpo es la más afectada, qué lugar es el más recurrente, qué actividad es la más frecuente, etc.

Incluso se pueden establecer relaciones entre variables tales como: qué parte del cuerpo es la más afectada en un puesto de trabajo determinado. En general, éste tipo de análisis bastaría y no resultaría necesario invertir esfuerzos en usar *data mining* para hallar patrones de accidentabilidad u otro tipo de conocimiento a partir de ciertas técnicas de minería de datos; sin embargo y tal como se ha descrito anteriormente, el *data mining* ofrece técnicas mucho más robustas que podrían detectar información oculta no muy perceptible en un estudio de frecuencias tradicional. La potencia del *data mining* reside en su capacidad de relacionar las variables a partir de modelos planteados por el analista.

No tiene caso revisar los principales modelos de gestión empresarial sólo advertir que es algo a tomar en cuenta o quizá lo principal, antes de usar esta tecnología. Por ejemplo, una empresa pequeña como una pyme quizá no lo requiera.

2.2. Obtención de conocimiento a partir de los datos de accidentabilidad laboral

Llegados a este punto se deben tener algunas cosas claras.

En primer lugar se debe tener el insumo principal del *data mining*: datos (ya se ha visto en las secciones anteriores como obtenerlos). Por otro lado, es preciso aclarar que el *data mining* ha sido diseñado para trabajar con grandes volúmenes de datos, aquellos que por su naturaleza son transaccionales. La accidentabilidad no es algo, al menos en ésta época, que genere muchos registros (recuérdese que cada registro es un caso o accidente) de modo que partiendo de allí podría resultar poco alentador usar esta herramienta; sin embargo sí podría ser interesante estudiar accidentes de distintos años o accidentes de distintos tipos. Es cuestión de criterio del analista.

Posteriormente se debe reflexionar sobre el para qué se quiere usar el *data mining*, es decir, qué es lo que se pretende encontrar. Este punto es clave, ya que sobre una correcta definición del problema y selección de herramientas de trabajo se puede tener alta probabilidad de obtener conocimiento fiable y real. En general el *data mining* se divide en tareas descriptivas y predictivas, cada una de ellas tiene una colección de técnicas sobre las cuales se derivan un sin número de métodos o algoritmos para resolver los modelos planteados por el analista. Por ejemplo, en el análisis predictivo se tiene la tarea de clasificación el cual puede resolverse usando redes neuronales. Esta información puede consultarse en capítulo 1.3.4 del presente trabajo. Ahora bien, el estudio de patrones de accidentabilidad es un modelo descriptivo – predictivo ya que intenta en el primer caso realizar un estudio descriptivo usando la tarea de agrupamiento para luego intentar hallar reglas de clasificación para futuros eventos. Es posible usar otras tareas y algoritmos asociados a estas tareas, según el analista lo requiera.

El tercer punto importante es la de herramienta o *software* a utilizar para diseñar el modelo de *data mining* y aplicar las tareas y técnicas seleccionadas por el analista. En el mercado existen muchas aplicaciones para realizar *data mining*, incluso se encuentran embebidas en muchos motores de gestión de bases datos tales como las herramientas *OLAP* del *SQL Server* de *Microsoft*. Un *software* muy potente en cuanto a la cantidad de algoritmos que contiene es el

*Weka*²², desarrollado por la Universidad de Waikato, Nueva Zelanda. Este *software* es de distribución gratuita aunque su interface no es muy amigable ni tampoco ofrece un motor de diseño gráfico del modelo del *data mining*.

Mencionados estos asuntos se procederá a detallar los pasos a seguir para hallar patrones de accidentabilidad laboral.

2.2.1. Aplicación del procedimiento estándar KDD

A saber, la aplicación de minería de datos tiene un espectro muy grande respecto a los estudios que podrían realizarse y a las técnicas que potencialmente podrían usarse en tales estudios en función de lo que se quiere estudiar por supuesto. Ciertamente cuando de intentar obtener conocimiento se trata, el uso de la metodología KDD descrito en el capítulo 1.3.3 del presente trabajo, resulta ser la más adecuada dado que es el método estándar. Sin embargo, este método no determina de ningún modo el modelo de *data mining* que se va a usar ya que esa responsabilidad recae en el analista. Lo cierto es que dicho procedimiento sirve para desarrollar y resolver el modelo.

El estudio de patrones de accidentabilidad busca encontrar características particulares, recurrentes y determinantes en la potencial ocurrencia de accidentes de trabajo a través de la aplicación de tareas descriptivas de *data mining*. De esta forma se puede describir con mayor detalle la dinámica de datos relevantes y registrados en la ocurrencia de accidentes laborales. Así también, se pretende hallar reglas de clasificación como tarea predictiva a fin de determinar con el mayor grado de certeza posible, las probabilidades de ocurrencia de accidentes de trabajo bajo estas reglas y el grado de congruencia de las mismas con los patrones de accidentabilidad hallados a fin de garantizar la solvencia del modelo de patrones.

Para llegar a realizar el *data mining* se sugiere utilizar como criterio restrictivo del modelo, que los datos seleccionados tengan relación causal con la ocurrencia de accidentes y cuya aparición en la secuencia del accidente sea obviamente antes de su ocurrencia o materialización. Esto resulta de considerar como referencia la teoría de causalidad de accidentes de trabajo. Se ha hablado al respecto en el capítulo 1.2.5 del presente trabajo.

Un ejemplo claro de lo anterior es considerar la hora de ocurrencia del accidente como variable de estudio mientras que el número de días de descanso médico que el trabajador tuvo por el accidente no, puesto que ese dato se registra como consecuencia del accidente. Naturalmente, los días de descanso médico pueden registrarse en la base de datos.

Figura 2.11. Obtención de patrones de accidentabilidad basado en el KDD

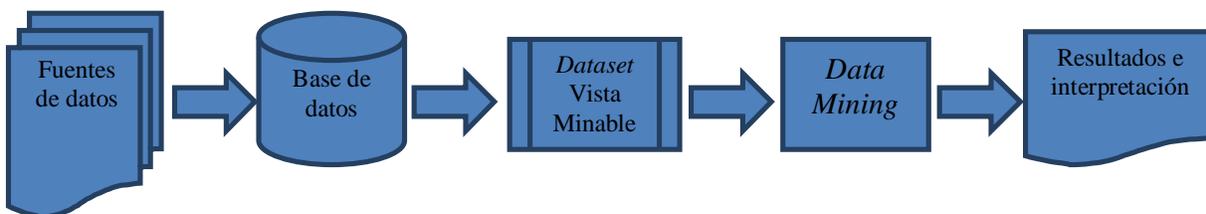


Diagrama de macro procesos del KDD aplicado a obtención de patrones de accidentabilidad

²²*Weka. The University of Waikato.* Disponible en <http://www.cs.waikato.ac.nz/~ml/weka/>

En términos generales se tiene lo siguiente:

1. Preparación del *dataset* o vista minable

Este primer paso es en resumidas cuentas, el trabajo que se realiza con los datos a fin de obtener una vista minable o *dataset* listo para aplicarle *data mining*. Este primer procedimiento incluye los pasos de selección, exploración, limpieza y transformación del *KDD*. Ya se había visto en la primera parte del presente capítulo los puntos más importantes respecto a este tema, pues se explica de manera puntual las variables implicadas en el diseño de una estructura de datos de accidentabilidad laboral y sus respectivas fuentes. Ahora bien, el modelo parte por definir una estructura de datos, la cual se ha sugerido sea definida en una única base de datos, para que a partir de allí se realicen los procesos de edición; es decir, exploración, limpieza y transformación antes de obtener el *dataset* final.

2. Modelado y desarrollo del *data mining*

En este punto ya se tiene la vista minable lista para aplicar algoritmos de *data mining* según el modelo que implemente el analista. Los modelos pueden ser descriptivos o predictivos y en función de ellos se aplican las tareas y algoritmos de minería de datos. Es importante utilizar las herramientas del *software* que se escoja para tener mayor espectro de opciones en cada paso del modelo pero sobre todo, es importante explotar las herramientas gráficas que puedan dar una mejor idea del comportamiento de los datos y los patrones que estos van formando.

3. Interpretación de resultados y conclusiones

Este paso es el último e igual de importante que los anteriores, puesto que en función de los resultados se deducen las conclusiones relacionadas al modelo planteado. En otras palabras, se obtienen los patrones que definen el comportamiento de la accidentabilidad de la empresa o institución analizada. Este paso confirmaría o rechazaría hipótesis *a priori*. Un ejemplo práctico de esto es determinar si efectivamente las fechas de pago a los trabajadores suponen un momento de alto riesgo de accidentabilidad para los mismos.

Vale acotar que el estudio de patrones de accidentabilidad describirán los patrones detectados sin explicar por qué ocurren o se asocian de tal o cual forma. Esa es tarea del analista y el grupo de expertos que participarían del experimento.

2.2.2.1. Selección, Exploración, limpieza y transformación.

a. Selección y exploración de datos.

Para el estudio de patrones de accidentabilidad se ha comentado que se deben considerar los siguientes criterios de selección:

Con respecto a las fuentes a fin de seleccionar campos que serán potenciales variables de estudio:

- Seleccionar campos de la ficha obligatoria de registro de accidentes de trabajo dispuesta por el Ministerio de Trabajo.
- Seleccionar campos particulares en función al contexto en que ocurren los accidentes de trabajo como por ejemplo: tipo de buque en la cual el tripulante sufrió el accidente de trabajo. Este caso aplicaría bien en la industria naviera.

Con respecto a los datos registrados en los campos seleccionados como potenciales variables de estudio.

- Aplicar técnicas de muestreo en caso se disponga de un universo de datos muy grande. Se ha mencionado anteriormente que no se pretende encontrar esta restricción en las bases de datos de accidentabilidad ya que la naturaleza de sus registros no es precisamente transaccional. Aun así y dependiendo del alcance de estudio, esta técnica podría resultar conveniente.

A momento se ha sugerido que los siguientes campos son potencialmente seleccionables como variables de estudio:

- Edad.
- Área.
- Puesto de trabajo.
- Antigüedad en el empleo.
- Sexo.
- Turno.
- Tipo de contrato.
- Tiempo de experiencia en el puesto de trabajo.
- Número horas trabajadas en la jornada laboral.
- Fecha y hora de ocurrencia del accidente de trabajo.

Los campos seleccionados deberían tener relación con causas potenciales de accidentes, éstos ocurren antes que el accidente se materialice; sin embargo y aunque no podría resultar recomendable, el analista podría “jugar” con los campos según su criterio. Quizás desee estudiar variables post accidente para ver su relación en los patrones de accidentabilidad. Por ejemplo, podría incluir el tipo de lesión determinada o la parte del cuerpo afectada, sabiendo a priori que esas variables no explican por qué ocurrió el accidente de trabajo, ya que son consecuencia del mismo. Aun así, el ensayo puede resultar bueno en técnicas descriptivas de *data mining*.

Los campos listados anteriormente han sido extraídos de la ficha de registro de accidentes de trabajo que exige la legislación peruana. Dichos campos tienen más o menos relación directa con la causalidad de los accidentes. El analista es libre de retirar campos si cree que tal relación no existe, así como agregar campos que según su criterio, sí tengan ésta relación. Los dos principales criterios de selección ya han sido comentados.

En este punto es muy importante tener en cuenta que para seleccionar los campos el analista deberá tener un conocimiento muy claro del programa de seguridad y salud en

el trabajo que está aplicando en la empresa o institución objeto de estudio. Por ejemplo, se sabe claramente que las competencias del personal son requisito indispensable para el desempeño del puesto de trabajo. Carecer de éstas no sólo predispone a un mal trabajo sino también a que ocurra un accidente laboral, por tanto las competencias del personal son variable directa en la causalidad de accidentes. El analista debe necesariamente consultar a personal experto.

En atención al punto anterior, obviar una variable de este tipo puede generar malas interpretaciones a pesar de que el modelo está bien planteado. Si por ejemplo, las competencias del personal no fueron incluidas en el estudio de patrones y resultó ser que la mayoría del personal accidentado carece de ellas, se llegarán a conclusiones que no impliquen necesariamente a esta variable.

Así mismo, podría ser que el analista obvие ciertas variables a priori dado que no tiene datos registrados asociadas a las mismas y que recomiende su registro para mejorar el estudio de patrones. Lógicamente al tener una visión más clara de los programas de seguridad industrial, podrá tener mejor panorama de análisis. Finalmente, es claro que no todas las instancias posibles del problema se van a tocar a la vez.

A estas alturas se tiene la base de datos definida y los campos que serían las potenciales variables asociadas a la ocurrencia de accidentes de trabajo. Por tanto lo siguiente a realizar es estudiar los mismos campos y sus respectivos datos a fin de garantizar que sean independientes y estadísticamente estables.

La exploración de datos puede iniciar con técnicas visuales sencillas a fin de verificar la distribución de los mismos. El uso de histogramas puede ser de gran utilidad.

Así mismo uno de los puntos más importantes en esta parte del proceso es tratar de reconocer los campos relacionados pues probablemente generen correlación estadística entre ellos. El ejemplo del lugar en donde ocurrió el accidente de trabajo refleja esta posibilidad. Si un trabajador se accidenta dentro de un taller que pertenece a un taller más grande el cual se ubica en una de las plantas de producción de toda la empresa, se tienen tres posibles lugares que se refieren con mayor a menor precisión al mismo lugar.

La correlación de variables no sería problema si es que no afectaran al desempeño de los algoritmos de base de datos de modo que lo más recomendable es dejar en el *dataset* únicamente variables independientes.

Así se puede tener el caso de la variable hora con el turno. Si por ejemplo una empresa cuenta con dos turnos: diurno y nocturno, no tendría caso agregar esa variable si vamos a considerar la variable “hora” en el *dataset*. Lo mismo ocurriría en el caso contrario.

El analista debería usar una tabla de correlación como herramienta más recomendable.

b. Limpieza y transformación.

En este punto se debe tener claro cuáles serán los campos seleccionados que se usarán como variables para el estudio de patrones de accidentabilidad. Se reitera que dichos campos deberán tener relación directa con las potenciales causas que pudieran originar accidentes, tales como la edad, el sexo, el tiempo de servicio, el turno u horario, el

lugar de trabajo, etc. Así como campos particulares según el trabajo que se realizaba tales como, el clima, temperatura o campos más especializados.

Así mismo ya se hicieron vistas exploratorias descriptivas para observar básicamente la distribución de los datos y obtener alguna idea sobre, por ejemplo, cuál es la edad más recurrente respecto a un accidente de trabajo. Entre otros temas importantes en la selección y exploración de datos según el *KDD* se encuentra la necesidad de hallar correlaciones entre variables. Es importante no obviar este paso a fin de optimizar el procedimiento para hallar patrones de accidentabilidad.

Teniendo más o menos establecida ya la vista minable, se debe realizar un análisis un poco más detallado a los datos individuales a fin de detectar valores atípicos (denominados *outliers* en la literatura de minería de datos) que puedan alterar de alguna forma la distribución general. Por ejemplo, se puede tener que la distribución de edades del personal de la plantilla de trabajadores oscila en un rango promedio de 25 a 55 años de edad, descubriéndose que en la data hay valores fuera de este rango. Estos valores son precisamente *outliers*. Ahora bien se debe tener especial cuidado con estos valores ya que puede ser que sean erróneos (probablemente el encargado de la base de datos cometió un error al momento de realizar su registro) o puede que sean valores legítimos solo que poco probables en la distribución general. Normalmente esto suele pasar con valores asociados a las características físicas de los trabajadores (peso, talla, índice de masa corporal, etc.). La metodología *KDD* ofrece varias alternativas para superar este tipo de problemas.

También es posible encontrar valores vacíos o *missing* en la base de datos. Es recomendable completar estos valores pues dejarlos vacíos podría perjudicar el desempeño correcto de los algoritmos asociados a tareas de *data mining*, sobre todo si la herramienta o *software* usado no prevé acciones para estos casos. En general se recomienda completar los valores con la media o moda del conjunto de datos. Quizá sea factible usar la mediana pero ello depende de la distribución.

Posteriormente se debe optimizar la dimensionalidad de la vista minable. Probablemente el analista quiera en primera instancia, utilizar todas las variables posibles para realizar el estudio de patrones, lo cual es bueno pero no muy recomendable ya que se puede caer en la popular **“maldición de la dimensionalidad”**²³ la cual se refiere a la posibilidad de obtener patrones poco fiables cuando los campos o variables son demasiados respecto a la cantidad de registros. Esto es particularmente importante en el estudio de patrones de accidentabilidad dado que se conoce a priori que muy probablemente no se tengan muchos registros. *KDD* ofrece técnicas para resolver estos problemas; sin embargo éstas pueden resultar ser muy complejas para el analista por lo cual es mucho mejor decidir la selección de campos en base a la experiencia y conocimiento que tenga sobre las características de los accidentes de trabajo que está estudiando.

Finalmente, un tema que no se puede obviar es la creación de características o nuevos campos a partir de los existentes. Esto a efectos de mejorar o adaptar los datos a los algoritmos de *data mining* simplemente aprovechar la información que aportan los datos individuales. Para entender mejor esta parte importante del trabajo previo que se

²³Hernández José, Ramírez José, Ferri Cesar (2004). *Introducción a la minería de datos*. Madrid: Pearson Prentice Hall. Pag. 79

hace con los datos es importante distinguir la categorización del dato o lo que es lo mismo, los tipos de datos (concepto clave en bases de datos). Normalmente se diferencian dos tipos de datos, aquellos que son numéricos y aquellos que son nominales o categóricos (texto, en términos sencillos). Naturalmente, las herramientas de *data mining* han desarrollado una vasta y compleja taxonomía de tipos de datos según sus propias necesidades, tales como el tipo fecha, hora, carácter, cadena y un largo etcétera.

Si por ejemplo se tiene el campo fecha en el siguiente formato: dd/mm/aaaa siendo “dd” día, “mm” mes y “aaaa” año es posible obtener tres campos distintos a partir del original. Uno para día, otro para el mes y finalmente otro para el año.

A propósito del tipo de dato es preciso referir que probablemente sea el aspecto más importante a tomar en cuenta antes de usar los algoritmos de *data mining*, ya que estos últimos están condicionados muchas veces al tipo de datos, de modo que algunos algoritmos están diseñados para trabajar con datos numéricos mientras que otros usan los datos categóricos o nominales. Un ejemplo sencillo de esto se aprecia en las medias y modas de la estadística descriptiva. La primera se aplica para valores numéricos, mientras que la segunda se usa en valores categóricos.

La mejor forma de resolver estos problemas técnicos es el uso de la discretización (conocida como *binning*) y la numerización de datos. En términos sencillos, los datos pasan de numéricos a categóricos y viceversa respectivamente.

En el estudio de patrones de accidentabilidad se pueden tener los siguientes casos:

- Edad: Este campo podría resultar conviene tratarlo como rangos de valores aplicando la técnica de discretización.
- Día, mes, año: Las variables temporales son cíclicas y ello puede generar problemas al momento de realizar la tarea de agrupamiento del *data mining* ya que el primer y último valor será representados como lejanos entre sí. Por ejemplo, el día lunes se encuentra cerca al martes pero lejos del sábado. La idea es que minimizar dicho impacto. Para resolver este problema las series deben ser ordenadas de forma distinta. A veces podría resultar mejor hacerlo de forma arbitraria.

2.2.2.2. Planteamiento del modelo de *data mining*

Construir el modelo del *data mining* es una tarea que se gesta desde que se empiezan a gestionar los datos con aras de estudiarlos a través de tareas de minería de datos. En términos generales, el modelo debe tener 03 componentes fundamentales:

1. *Input*.-Bloque de entrada de datos al modelo conformado básicamente por la vista minable, en la cual sus datos han sido previamente, seleccionados, muestreados, transformados, etc.
2. *Data Mining*.- De acuerdo a la tarea que se desea ejecutar se tienen técnicas predictivas y descriptivas. El analista debe escoger los algoritmos de minería de datos según sea el caso. Por ejemplo se puede tener un modelo de minería de datos basado en un árbol de decisión.

3. *Output*.- Se obtienen resultados del modelo aplicado a los datos a fin de que sean interpretados.

Para el estudio de patrones de accidentabilidad se realizará un modelo descriptivo usando la tarea de agrupamiento a través del algoritmo K-means. Posteriormente se valida y se obtiene a la vez un modelo predictivo a través de la tarea de clasificación basada en el algoritmo C 4.5.

El experimento a realizar estará conformado por los siguientes pasos:

1. Aplicar el algoritmo *K-means* en la tarea de *clustering* (agrupamiento) para encontrar patrones comunes en las variables de estudio.
2. Interpretar los resultados obtenidos haciendo uso de las herramientas que ofrece el *software* de *data mining* bajo la opinión de expertos.
3. Aplicar el algoritmo de inducción C4.5 en la tarea de clasificación para identificar reglas de pertenencias de un valor a cada grupo.
4. Obtener conclusiones finales.

En cuanto al método que se va a realizar, se propone el siguiente procedimiento:

Tabla 2.1. **Procedimiento a seguir para obtener patrones de accidentabilidad**

Pasos KDD	Pasos según problema específico
Selección, Exploración, limpieza y transformación.	1. Obtener un <i>dataset</i> usando como fuente la base de datos de accidentabilidad.
Minería de datos (uso de técnicas descriptivas).	2. Aplicar el algoritmo <i>K-means</i> para encontrar grupos con características similares en función de los datos.
Evaluación e interpretación de resultados.	3. Validar e interpretar resultados con personal experto.
Minería de datos (uso de técnicas predictivas).	4. Aplicar el algoritmo de inducción C4.5 para obtener reglas de pertenencia a cada uno de los grupos obtenidos.
Evaluación e interpretación de resultados.	5. Con ayuda de expertos obtener conclusiones definitivas y generar recomendaciones.

Fuente: Adaptación del autor.

Respecto a los algoritmos de minería de datos a usar se tiene la siguiente descripción.

1. Algoritmo K-Means.

Este algoritmo se encuentra dentro del grupo de técnicas de minería de datos descriptiva. Su objetivo es agrupar los datos en un conjunto de K grupos distintos en función del centroide de cada grupo. A saber, el centroide representa el centro de cada grupo o *cluster* en función de la distancia euclídea que existe entre cada una de sus instancias. Estos cálculos se hacen a través de la media (*mean*) o la moda (*mode*) según el tipo de dato con el cual se esté trabajando, el cuál puede ser numérico o categórico, respectivamente.

A continuación se describe el pseudocódigo del algoritmo *K-means*²⁴:

1. Elegir k ejemplos que actúan como semillas (k número de *clusters*).
2. Para cada ejemplo, añadir ejemplo a la clase más similar.
3. Calcular el centroide de cada clase, que pasan a ser la nueva semilla.
4. Si no se llega a un criterio de convergencia (por ejemplo, dos iteraciones no cambian las clasificaciones de los ejemplos), volver a 2.

Los ejemplos representan las instancias de cada grupo y estos se van asignando a cada grupo o clase en función de sus distancias euclídeas con respecto al centroide, de modo que se asegure la minimización de estas distancias (la varianza total intra grupo debe ser la menor posible). Así mismo, en cada iteración se va recalculando el centroide y se valida si es que se ha llegado a la convergencia, momento en el cual no hay datos que cambien de grupo de una iteración a otra. En este punto, el algoritmo ha terminado.

No obstante y puesto que el algoritmo se basa en un inicio arbitrario no hay garantía de obtener la mejor posible asignación en la primera prueba, por lo que se recomienda ejecutar varias veces el algoritmo con distintas semillas.

2. Algoritmo de inducción C4.5

El algoritmo C4.5 pertenece a la familia de los *Top Down Induction Trees (TDIDT)* los cuales son algoritmos de aprendizaje automático a partir de ejemplos predeterminados. En minería de datos se utiliza para las tareas de clasificación de datos en árboles de decisión.

Los sistemas de aprendizaje basados en árboles de decisión son los más populares y probablemente sencillos del espectro de algoritmos de aprendizaje. Éstos se estructuran de manera jerárquica partiendo de un nodo raíz hasta sus algunas de hojas a través de condiciones organizadas y establecidas. Las opciones posibles que se deducen del árbol de decisión resultan excluyentes entre ellas lo cual proporciona una gran ventaja al momento de la toma de decisiones pues solo es posible una opción de entre las demás.

Los algoritmos usados para construir árboles de decisión se llaman de cobertura o de “divide y vencerás” y deben considerar los siguientes puntos para que su funcionamiento sea óptimo²⁵:

- Particiones a considerar.
- Criterio de selección de particiones.

ID3 es un algoritmo simple y, sin embargo, potente, cuya misión es la elaboración de un árbol de decisión. El procedimiento consiste básicamente en lo siguiente: Seleccionar un atributo como raíz del árbol y crear una rama con cada uno de los posibles valores de dicho atributo. Con cada rama resultante (nuevo nodo del árbol), se realiza el mismo proceso, esto es, se

²⁴ Molina José, García Jesús (2006). *Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y Weka*. Univeridad Carlos III de Madrid. Pag. 99.

²⁵ Hernández José, Ramírez José, Ferri Cesar (2004). *Introducción a la minería de datos*. Madrid: Pearson Prentice Hall. Pag. 283-284.

selecciona otro atributo y se genera una nueva rama para cada posible valor del atributo²⁶. El nodo hoja representa la clase o etiqueta que clasifica al ejemplo o instancia.

Con respecto al criterio de selección de particiones de ID3 se usa el criterio *Gain* (ganancia) el cual se basa en el concepto de entropía. La idea es entonces encontrar el parámetro o atributo más significativo, aquel que permita discriminar mejor los ejemplos o instancias. ID3 sólo funciona bien para variables discretas precisamente porque el criterio de selección de particiones se basa en la entropía lo cual haría ineficiente computacionalmente el rendimiento del algoritmo al tomar cada uno de los posibles valores del atributo continuo.

C4.5 es una mejora del ID3 el cual fue creado por Ross Quinlan. C4.5 también pertenece al mismo autor. Entre sus principales características destacan²⁷:

- Permite trabajar con valores continuos para los atributos, separando los posibles resultados en 2 ramas $A_i \leq N$ y $A_i > N$.
- Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases no una clase en particular.
- Utiliza el método “divide y vencerás” para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
- Se basa en la utilización del criterio de proporción de ganancia (*gain ratio*), definido como $I(X_i, C)/H(X_i)$. De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas con la selección
- Es recursivo.

El pseudocódigo del algoritmo C4.5 es el que se muestra en la figura 2.11.

2.2.2.3. Evaluación e interpretación de resultados

Aplicado el algoritmo *K Means* el analista obtendrá como resultado el agrupamiento de datos según los *clusters* determinados con un porcentaje de distribución para cada caso. Es importante comprobar que todas las instancias que se han considerado en el modelo hayan sido utilizadas por el algoritmo, caso contrario se tendría que revisar nuevamente el modelo a fin de validar que los tipos de datos y las transformaciones sean las más adecuadas.

Es importante destacar los resultados más representativos dado que dan una visión más amplia de los patrones encontrados. El *software Weka* permite obtener estos resultados luego de que se ha corrido el algoritmo.

Posteriormente se aplica el algoritmo C 4.5. Lo primero a realizar es la selección de atributos. Para esto se puede utilizar como referencia los métodos empleados por Ignacio Perversi²⁸. La idea en este punto es encontrar el subconjunto de atributos que permita explicar un atributo objetivo. Se aplica una tarea de clasificación supervisada. Lo que se quiere es dado los

²⁶ Molina José, García Jesús (2006). *Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y Weka*. Univeridad Carlos III de Madrid. Pag. 125.

²⁷ López Takeyas Bruno. Algoritmo C4.5. En línea Internet. Noviembre de 2005. Accesible en [http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5\(2005-II-B\).pdf](http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf)

²⁸ Perversi, Ignacio (2007). *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina*. Tesis de grado en ingeniería industrial. Buenos Aires: Instituto Tecnológico de Buenos Aires. Pag. 68.

clusters obtener la serie de atributos cuyas instancias interrelacionadas permitan obtener cada uno de ellos.

En la tabla descrita a continuación se describen los algoritmos y métodos más utilizados para detectar los atributos más “influyentes” en el modelo descriptivo.

Figura 2.12. Pseudocódigo del algoritmo C4.5

<p>Función C4.5 R: conjunto de atributos no clasificadores, C: atributo clasificador, S: conjunto de entrenamiento, devuelve un árbol de decisión Comienzo Si S está vacío, Devolver un único nodo con Valor Falla; ‘para formar el nodo raíz Si todos los registros de S tienen el mismo valor para el atributo Clasificador, Devolver un único nodo con dicho valor; un único nodo para todos Si R está vacío, Devolver un único nodo con el valor más frecuente del atributo Clasificador en los registros de S [Nota: habrá errores, es decir, Registros que no estarán bien clasificados en este caso]; Si R no está vacío, D ← atributo con mayor Proporción de Ganancia (D,S) entre los atributos de R; Sean {dj j=1,2,..., m} los valores del atributo D; Sean {Sj j=1,2,..., m} los subconjuntos de S correspondientes a los valores de dj respectivamente; Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1, d2,...,dm, que van respectivamente a los árboles C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm); Fin</p>
--

Tabla 2.2. Selección de atributos.

Algoritmo de evaluación	<i>Wrapper</i>	<i>Classifier</i>	<i>Consistency</i>	<i>CfsSubsetEval</i>	<i>ChiSquare</i>	<i>Gain Ratio</i>
Método de búsqueda	<i>Exahustive</i>	<i>Exahustive</i>	<i>Exahustive</i>	<i>Exahustive</i>	<i>Ranker</i>	<i>Ranker</i>

Fuente: El autor.

Seleccionados los atributos más relevantes, se procede a correr el algoritmo C4.5 para lo cual se pueden combinar distintos grupos de atributos según su mayor a menor influencia al modelo descriptivo.

Finalmente se debe evaluar la matriz de confusión de cada combinación a fin de determinar el grado de precisión que tuvo el modelo para clasificar las instancias en los *cluster* hallados. Lo ideal es que la precisión del modelo sea del 100%, lo cual quiere decir que todas las instancias lograron ser clasificadas y por tanto se puede esperar que futuras instancias tengan un comportamiento similar con lo cual se tiene un modelo predictivo robusto.

²⁹López Takeyas Bruno. Algoritmo C4.5. En línea Internet. Noviembre de 2005. Accesible en [http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5\(2005-II-B\).pdf](http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf)

2.3. Aplicación del conocimiento obtenido al programa de Seguridad y Salud en el Trabajo

Esta es la parte objetivamente más importante de la metodología pues determina las decisiones que la empresa o institución ejecutaran a fin de retroalimentar su sistema de gestión de seguridad y salud en el trabajo. Las empresas trabajan en base a resultados y lo que esperan es tomar decisiones claras y medibles respecto a la gestión de la seguridad y salud ocupacional de su personal de modo que las conclusiones no pueden ser ambiguas y el conocimiento obtenido debe ser aprovechado de la mejor forma.

Es posible que las conclusiones impliquen temas que no se encuentren dentro del alcance del encargado del sistema de gestión. Un ejemplo clásico de esto es la definición de las paradas técnicas de trabajo para evitar accidentes o enfermedades profesionales relacionadas a ergonomía. Normalmente el trato laboral a pesar que está asociado a las medidas de seguridad para evitar accidentes, no necesariamente van de la mano, incluso pueden ser incompatibles.

Por otro lado el analista debe reconocer que hay muchas variables que probablemente no fueron tomadas en cuenta y que potencialmente podrían tener una influencia considerable en el modelo global. El analista debe tomar estas posibilidades como oportunidades de investigación y mejora de la base de datos.

Los programas de seguridad y salud en el trabajo tienen una distribución básicamente como sigue:

1. Análisis de riesgos y agentes.
2. Programas de entrenamiento y capacitación.
3. Programas de monitoreo de factores de riesgo.
4. Diseño de estándares o procedimientos.
5. Ejercicios de entrenamiento y respuesta a emergencias.
6. Gestión del personal.
7. Investigación de accidentes.

El conocimiento obtenido será de mucha ayuda en soluciones o propuestas asociadas a los programas previos a la ocurrencia de accidentes. Desde el análisis de riesgo hasta la investigación de accidentes, se presentará un *input* interesante de parámetros o datos a tener en cuenta.

Un breve resumen de las posibilidades de mejora que se tienen, se presentan a continuación:

1. Mejora de los turnos de trabajo.
2. Mejora de los programas de capacitación y competencias del personal.
3. Oportunidades de estudio e ingeniería para mejora de las condiciones de los centros de trabajo. Esto se aprecia muy bien cuando se obtienen patrones marcados de accidentes en ciertos lugares de trabajo.
4. Mejora de la gestión de personal respecto a su edad, condiciones físicas y su asignación a ciertas locaciones de trabajo.
5. Oportunidades de estudio y mejora de los EPP más adecuado a los cargos más frecuentes de accidentes o según su locación de trabajo.
6. Mejora en los procedimientos de trabajo que podrían tener relación directa con la aparición de accidentes.
7. Mejora en los procedimientos de investigación de accidentes.
8. Mejora en la respuesta a condiciones estacionales como lo es clima.

Se puede apreciar, las posibilidades de acción a partir del conocimiento obtenido, son muy grandes. El analista y el equipo experto deben tener una visión muy amplia para lograr soluciones objetivas y eficaces.

Capítulo 3

Conclusiones y recomendaciones

La metodología expuesta en el presente documento no pretende implantar un método único e infalible para determinar patrones de accidentabilidad. El espectro de investigación y las posibilidades que ofrece el *data mining* son muy amplias como para pretender establecer un estándar único. No obstante, recomienda una serie de pasos a seguir para obtener un resultado específico que es hallar patrones de accidentabilidad. Ya se ha mencionado que las empresas tienen la posibilidad de usar la estadística descriptiva tradicional con la salvedad que la posibilidad de obtener conocimiento es reducida. El *data mining* ofrece herramientas más robustas para este fin.

Por otro lado, es preciso reconocer que actualmente el Perú vive un contexto nuevo en relación a la gestión de la seguridad y salud de sus trabajadores. La publicación de la legislación en estas materias ha establecido un nuevo orden de trabajo para las empresas en las que esta disciplina ya no será considerada como “algo más” sino que pasa a ser protagonista con una concepción más interesante: ya no se trata de un gasto, sino de una inversión. El conocimiento que se pueda obtener para entender la dinámica de la accidentabilidad de las empresas resulta crucial en la implementación de medidas claras.

Este trabajo debe ser tomado como una guía referencial ya que en cierta medida es abstracto respecto a los caminos que debe seguir un analista, es imposible determinar lo que se debe hacer en cada etapa del proceso respecto al análisis que cualquiera decida realizar, aunque sí se pueden dar recomendaciones y pautas a seguir, siempre apoyados en la metodología *KDD*.

Este estudio académico concluye con la descripción de la aplicación real de la metodología a una empresa real a fin que se pueda obtener una visión más gráfica de lo que implica realmente aplicar la metodología a cualquier realidad. Es importante recordar que el espíritu del trabajo es ayudar a cualquier analista a diseñar y aplicar modelos de *data mining* para estudiar la accidentabilidad en su sector de estudio, la idea es que encuentre patrones característicos de su propia realidad.

A continuación se proponen algunas líneas de investigación asociadas a este tema.

Línea de investigación en análisis de riesgos

La legislación pide que las empresas analicen sus riesgos para implementar medidas de control sobre los mismos. La metodología IPER (Identificación de peligros y evaluación de los riesgos) es un método que permite evaluar y ponderar los riesgos derivados de cada peligro identificado en tareas de una determinada actividad. Así se pueden agrupar las tareas con riesgos más significativos, los cuales se obtienen de acuerdo a ciertos parámetros predeterminados, a fin de poder implementar medidas de control específicas que minimicen lo más que se pueda al riesgo tratado.

La aplicación de minería de datos en este aspecto permitiría estudiar que dadas las condiciones de control planteadas en el IPER se podría registrar la ocurrencia o no del accidente a fin de evaluar si estas medidas de control resultan factibles o si se está obviando alguna(s) otra(s) variable(s). Así mismo y a nivel predictivo un objetivo supremo sería obtener probabilidades más confiables de ocurrencia de accidentes dadas ciertas condiciones iniciales, esto a través de un modelo de clasificación por ejemplo.

A medida que se obtienen más datos de trabajos en las cuales dadas ciertas medidas de control se registran la ocurrencia de accidentes en las cuales puede haber o no se permitirá encontrar mejores patrones de comportamiento de los mismos para casos especiales y sobre todo para descifrar variables de control no consideradas, es interesante que se pueda tomar en cuenta el estudio del riesgo oculto usando estas técnicas.

Línea de investigación en gestión de accidentes

La minería de datos está diseñada en esencia para trabajar con grandes volúmenes de datos y variables. Del estudio de variables específicas en el presente trabajo se deduce que la potencia en el hallazgo de patrones será igual de efectivo para el estudio de casos muy particulares que involucran variables más específicas. Por ejemplo, el estudio de accidentes en barcasas de trabajo considerando como posible variable de estudio el día del turno con respecto al final o comienzo de éste en el que el trabajador resulto accidentado resultaría interesante para tratar de comprender hasta qué punto hay una relación entre los accidentes y el día en que ocurren a personal recién integrado a su turno o a punto de salir (esto es evidentemente útil para el personal embarcado que por ejemplo tiene que laborar 14 días a bordo).

Finalmente se puede considerar una serie de posibilidades a explotar con esta herramienta, las cuales con la experiencia en su uso y la estandarización de los métodos de análisis de accidentes puede generar oportunidades de mejora en los métodos de prevención que tanto se necesita en los sectores industriales.

Bibliografía

Abdul Raouf. Teoría de las causas de accidentes. Pag. 56.6 – 56.7. Disponible en <http://www.insht.es/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

Asfahl, C. Ray. (2000). *Seguridad industrial y salud*. 4º ed. Prentice Hall. Pag. 4.

Banco Mundial (2010). *El mercado laboral peruano durante el auge y caída*. Perú: Ledel SAC. Pag. 27.

Bestraten Belloví Manuel, Gil Fisa Antonio, Pique Ardanuy Tomás. *NTP 592: La gestión integral de los accidentes de trabajo (I): tratamiento documental e investigación de accidentes*. España. Ministerio de Trabajo y Asuntos Sociales. Instituto Nacional de Higiene y Seguridad en el Trabajo. Disponible en: http://www.insht.es/InshtWeb/Contenidos/Documentacion/FichasTecnicas/NTP/Ficheros/501a600/ntp_592.pdf

Cortez Díaz, José. (2007). *Seguridad e higiene del trabajo. Técnicas de prevención de riesgos laborales* 9º ed. Tebar. Pag. 43.

Fraile Cantalejo Alejo, Lopez Beneyto Felix, Maqueda Blasco Jerónimo, Muñoz Muñoz Angel, Obregón Cagiga Pedro, Pique Ardanuy Tomás, Rosel Ajamil Luis, (1993). *Proyecto INVAC: una contribución a la modernización de la investigación de Accidentes de trabajo*. Disponible en: http://www.mapfre.com/documentacion/publico/i18n/catalogo_imagenes/grupo.cmd?path=1014047

Fundación Universitaria Iberoamericana FUNIBER. *Business Intelligence y Gestión Documental*. Sección. 1.3.2.2.

García Morate, Diego. Manual de Weka. Disponible en <http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf>

Hernández Zuñiga Alfonso, (2005). *Seguridad e higiene industrial*. Mexico: Limusa. Pag. 29 – 32, 79.

Hernández José, Ramírez José, Ferri Cesar (2004). *Introducción a la minería de datos*. Madrid: Pearson Prentice Hall. Pag. 05 – 06, 17, 20, 148, 283 – 284.

IMI del Perú SAC, (2012). *Procedimiento para investigación de accidentes P-901707*.

IMI del Perú SAC. *Plan estratégico 2010 - 2016*.

López Takeyas Bruno. Algoritmo C4.5. En línea Internet. Noviembre de 2005. Disponible en [http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5\(2005-II-B\).pdf](http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf)

Ministerio del Trabajo. Anuario Estadístico. Pag. 56.2. Disponible en <http://www.mintra.gob.pe/mostrarContenido.php?id=86&tip=87>

Ministerio de trabajo. Protocolo para la investigación de accidentes de trabajo. Disponible en http://www.mintra.gob.pe/archivos/file/dnit/protocolo_investigacion_accidentes_trabajo.pdf

Ministerio de trabajo. Sistema de accidentes de trabajo. Disponible en <https://www.sunat.gob.pe/xssecurity/SignOnVerification.htm?signonForwardAction=http://luna.mintra.gob.pe/si.sat/index.jsp>

Molina José, García Jesús (2006). *Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y Weka*. Univeridad Carlos III de Madrid. Pag. 99, 125.

Pérez César, Santín Daniel (2007). *Minería de Datos Técnicas y Herramientas*. Ed. 01. Madrid: Ediciones Paraninfo S.A. Pag. 06.

Perversi, Ignacio (2007). *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina*. Tesis de grado en ingeniería industrial. Buenos Aires: Instituto Tecnológico de Buenos Aires.

Piqué Ardanuy Tomás (1997). *Investigación de accidentes – incidentes. Procedimiento*. España. Ministerio de Trabajo y Asuntos Sociales. Instituto Nacional de Higiene y Seguridad en el Trabajo. Pág. 1 – 2.

Presidencia de la República del Perú. (2012). *Reglamento de la Ley N° 29783 Seguridad y Salud en el Trabajo*. Glosario de términos. Artículo 42, literal 1°. Artículo 65.

Saari, Jorma. Accidentes y Gestión de la Seguridad. Enciclopedia de Salud y Seguridad en el Trabajo de la OIT. Pag. 56.2, 56.3. Disponible en <http://www.insht.es/0057F17F-E555-4734-B4E9-5020B0978F88/FinalDownload/DownloadId-E81887714E078E9BA83EDE7E6D79BEA1/0057F17F-E555-4734-B4E9-5020B0978F88/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo2/56.pdf>

Weka. *The University of Waikato*. Disponible en <http://www.cs.waikato.ac.nz/~ml/weka/>

Vieira Luis, Ortiz Luis, Ramírez Santiago (2009). *Introducción a la minería de datos*. Brasil: E-papers Servicios Editoriales. Pag. 20.

Anexos

Anexo A - Listado de Tablas y gráficos

a. Tablas

- ✓ Tabla 1.1. Resumen de estadísticos del anuario 2010 Mintra (accidentes de trabajos por meses).
- ✓ Tabla 1.2. Técnicas de minería de datos.
- ✓ Tabla 2.1. Procedimiento a seguir para obtener patrones de accidentabilidad.
- ✓ Tabla 2.2. Selección de atributos.

b. Figuras

- ✓ Figura 1.1. Relación entre peligro, riesgo y consecuencia.
- ✓ Figura 1.2. Estructura de los accidentes.
- ✓ Figura 1.3. Arquitectura de un sistema de inteligencia de negocio basado en datos estructurados.
- ✓ Figura 2.1. Esquema de la metodología D.O.A
- ✓ Figura 2.2. Registro y notificación de accidentes de trabajo.
- ✓ Figura 2.3. Estructura de datos de accidentabilidad.
- ✓ Figura 2.4. Datos referidos al empleador.
- ✓ Figura 2.5. Datos referidos al trabajador.
- ✓ Figura 2.6. Datos referidos a la investigación del accidente de trabajo.
- ✓ Figura 2.7. Datos referidos a la descripción de hechos del accidente de trabajo.
- ✓ Figura 2.8. Datos referidos a las causas del accidente de trabajo.
- ✓ Figura 2.9. Medidas correctivas del accidente de trabajo.
- ✓ Figura 2.10. Proceso de investigación de accidentes de trabajo.
- ✓ Figura 2.11. Obtención de patrones de accidentabilidad basado en el KDD.
- ✓ Figura 2.12. Pseudocódigo del algoritmo C4.5.

Anexo B- Cuadro detallado de la ley SST respecto de la gestión de accidentes.

NORMATIVIDAD NACIONAL CRITERIO	Ley 29783	OBSERVACIÓN	D.S. 005-2012-TR	OBSERVACIÓN
MEJORA CONTINUA	Artículo 46.	La investigación de accidentes es información válida para tomar disposiciones con respecto al mejoramiento continuo.	Artículo 83.	El empleador debe adoptar disposiciones necesarias en materia de prevención, ante accidentes de trabajo.
CONSECUENCIA DE ACCIDENTES	Artículo 53.	El incumplimiento del empleador del deber de prevención genera la obligación de pagar las indemnizaciones a las víctimas de los accidentes de trabajo.		
	Artículo 76	Derecho del trabajador de ser transferidos sin menoscabos de sus derechos remunerativos.		
INFORMACIÓN DE ACCIDENTES DE TRABAJO	Artículo 80. Artículo 81.	Beneficio de los registros, notificaciones e investigación de accidentes para el Ministerio de Trabajo y Promoción del Empleo.	Artículo 117.	Ministerio de Trabajo y Promoción del Empleo es el responsable de examinar la información en materia de registro y notificación de los accidentes de trabajo
REPORTE DE ACCIDENTES	Artículo 82.	Deber de los empleadores informar ante el sector trabajo todo tipo de accidentes e incidentes.	Artículo 110. Artículo 122. Artículo 111. Artículo 121. Artículo 112. Artículo 113.	Plazo máximo para reporte de accidentes de trabajo. Empleo de sistema informático para reporte de accidentes. Notificación de accidentes por formularios físicos.
	Artículo 83.	Deber del empleador, que contrate labores bajo tercerización, de reportar la ocurrencia de accidentes.	Artículo 116.	Obligación de la comunicación de accidentes de contratistas.
REGISTROS DE ACCIDENTES	Artículo 87.	Obligación de contar con registro de accidentes de trabajo e incidentes peligrosos.	Artículo 33.	Registros obligatorios del Sistema de Gestión de Seguridad y Salud en el Trabajo.
	Artículo 88.	Exhibición y archivo de registros.	Artículo 35.	Periodo que deben conservarse los registros de accidentes.
INVESTIGACIÓN DE ACCIDENTES DE TRABAJO	Artículo 92.	Responsables de realizar la investigación de accidentes, y la obligación de comunicarlo a la autoridad administrativa del trabajo.	Artículo 42. Artículo 65.	Funciones del Comité de Seguridad y Salud en el Trabajo.
	Artículo 58	Cuando se debe realizar un investigación de accidentes.	Artículo 119.	La dirección de inspección del trabajo prioriza la inmediata fiscalización de los accidentes de trabajo mortales.
	Artículo 93.	Finalidad de las investigaciones de accidentes.		
	Artículo 94.	La autoridad administrativa de trabajo publica informes que entrañen situaciones de grave riesgo efectivo o potencial para los trabajadores o la población.		

Anexo C- Aplicación real del modelo de obtención de conocimiento

1. La Empresa

IMI DEL PERÚ S.A.C. (IMI), es una empresa dedicada a la prestación de servicios a la Industria Petrolera en el rubro de transporte marítimo, operación de embarcaciones y otras actividades conexas.

Sus oficinas principales se encuentran ubicadas en el Edificio Petro-Perú S/N, frente a los Parques 33-34 Playa Tortuga, Talara.

En febrero de 2000, la empresa IMI empieza sus operaciones ofreciendo servicios de transporte marítimo, reparaciones y mantenimiento de plataformas petroleras en el mar del norte del Perú, quien posteriormente incrementa sus servicios.

Actualmente brinda los siguientes servicios principales dentro del centro de operaciones:

1. Servicio de operación de barcos.
2. Servicio de operación de barcasas.
3. Servicio de gasfitería industrial en el mar.
4. Servicio de arenado y pintura.
5. Servicio de construcción y reparación de plataformas petroleras.
6. Servicio de metal mecánica.

2. Solución propuesta al problema específico

Al respecto, se propone aplicar tecnología de minería de datos sobre la información registrada en la base de datos de accidentabilidad de IMI. Para ello se propone el uso del software gratuito *Weka 3.6*. *Weka* es un *software* de distribución libre y gratuita, así mismo es una herramienta para realizar minería de datos muy potente ya que cuenta con un vasto set de algoritmos a usar.

2.1. Fuentes de información para el análisis.

Lo primero a realizar en el proceso de descubrimiento de conocimiento es la selección de los datos. Actualmente IMI no cuenta con una base de datos relacional para registrar los distintos accidentes que ocurren en sus operaciones. Los hechos se almacenan en una hoja electrónica *Microsoft Excel*. De este archivo se deberá obtener el *dataset* que servirá para realizar el análisis de minería de datos.

El *dataset* deberá contar con los campos o atributos que se pretenden analizar así como la depuración de registros con datos válidos para mejor lectura del *Weka*.

La base de datos contiene información referente a la descripción del accidente; esto es, las circunstancias en las cuales ocurrió: lugar, hora, fecha, trabajo a realizar, etc.; las características del incidente: tipo de lesión, parte de cuerpo afectada; datos del trabajador o trabajadores afectados y registro de causas inmediatas – básicas del evento.

El modelo descriptivo deberá estar basado en atributos que definan características relevantes antes que ocurra el hecho, pues a partir de estos datos se analizarán los patrones que describen las características de los accidentes registrados en IMI. Ciertamente, un accidente está

conformado por una serie de eventos y factores relacionados entre si, normalmente bajo la forma causa – efecto; sin embargo todas estas ocurren en un contexto que, a saber, se rige por características de tipo temporales (medidas de tiempo en las que ocurrió el accidente), espaciales (medidas de lugar donde ocurrió el accidente), personales (características de la persona que sufrió el accidente) y ambientales (características del entorno: clima, equipos, herramientas, etc.).

Para el proceso de exploración de datos del proceso de descubrimiento del conocimiento se ha obtenido la opinión del personal que trabaja en el Departamento de QHSE de IMI en consideración a los criterios de contexto explicados anteriormente. Sobre esta tabla resultante se aplicarán técnicas de limpieza y transformación del proceso de descubrimiento de conocimiento. La descripción de atributos de interés contendrá los siguientes parámetros: Nombre, descripción general y estados posibles.

Vale acotar que la denominación variable, atributo o campo representan el mismo significado para efectos del presente estudio.

2.2. Atributos de interés:

Tabla 2.1. Descripción de atributos seleccionados

Nombre	Descripción General	Estados Posibles
Etapa de vida	Representa el intervalo de tiempo respecto a la edad del trabajador que sufrió el accidente.	Entre 18 y 70 años de edad.
Cargo	Es el cargo en la empresa que desempeña el trabajador accidentado.	Anexo C.2.
Experiencia	Es la cantidad de años que lleva laborando el trabajador en la empresa.	Nuevo (personal que lleva laborando menos de 06 meses en el mismo puesto de trabajo).
		Intermedio (personal que lleva laborando en el mismo puesto entre 06 meses a 01 año).
		Antiguo (personal que labora en el mismo puesto de trabajo más de 01 año).
		Estos criterios han sido definidos por la empresa.
Día	Representa el día de la semana en que ocurrió el hecho.	Lunes
		Martes
		Miércoles
		Jueves
		Viernes
		Sábado
		Domingo
Mes	Representa el mes del año en la cual se accidento el trabajador.	Enero
		Febrero
		Marzo
		Abril
		Mayo
		Junio
		Julio
		Agosto
		Setiembre
		Octubre
		Noviembre
		Diciembre

Turno	Representa el intervalo de tiempo de acuerdo a la jornada laboral en que ocurrió el hecho.	Mañana (0601 – 1200 hrs.)
		Tarde (12:01 – 1900 hrs.)
		Noche (1901 – 0000 hrs.)
		Madrugada (0001 – 0600 hrs.)
Lugar	Representa el lugar o espacio físico donde ocurrió el hecho.	Anexo C.3
Tierra-Mar	Representa la zona de trabajo macro donde ocurrió el hecho.	Tierra
		Mar
Dotación-Servicio	Indica si el trabajador accidentado era propio del área donde ocurrió el accidente o se trataba de un tercero.	Dotación
		Servicio
Tipo de Trabajo	Indica si el trabajo a realizar era de rutina, especial o si por el contrario el accidente ocurre en circunstancias en que el trabajador no realizaba ningún trabajo.	Rutinario
		Especial
		No aplica (NA)

Fuente: El autor.

2.3. Atributos no considerados

- N°: representa el número de registro que corresponde a un accidente. No se considera pues no aporta información descriptiva del hecho.
- Año: representa el año en que ocurrió el accidente. Se ha obviado pues el estudio ha considerado un único periodo de tiempo el cual abarca del año 2005 a 2011.
- Tipo de incidente: representa el tipo de incidencia según lo establecido por la ley.
- Trabajador: contiene los nombres y apellidos del trabajador. No se considera pues no aporta información descriptiva relativa al hecho.
- Departamento: representa el nombre de la unidad organizacional a la cual pertenece el trabajador según la distribución descrita en el organigrama. No se considera como atributo relevante pues se ha escogido el atributo “cargo” del cual se puede obtener el área de trabajo a la cual pertenece.
- Dónde: representa el lugar exacto donde ocurrió el hecho. No es considerado dado que es un campo con mucho nivel de detalle. Se ha optado por escoger el campo “Lugar” el cual contiene las categorías de los sitios donde se realizan los trabajos.
- Causa inmediata: representa la causa directa que generó el accidente. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.
- Naturaleza: representa el riesgo o los riesgos asociados al accidente. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.
- Tipo de lesión: representa el tipo de lesión provocada por el accidente. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.
- Parte del cuerpo afectada: contiene la parte del cuerpo lesionada según distribución anatómica estándar. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.
- Diagnóstico: representa el diagnóstico clínico asociado a la lesión del trabajador. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.

- **Severidad:** representa los días de descanso médico necesarios para una total recuperación por parte del trabajador. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.
- **Causas básicas:** Representan las causas fundamentales que originaron el accidente. No se considera pues los valores de este atributo se registran después que ocurre el accidente mientras que lo que se quiere analizar son las características previas al hecho.

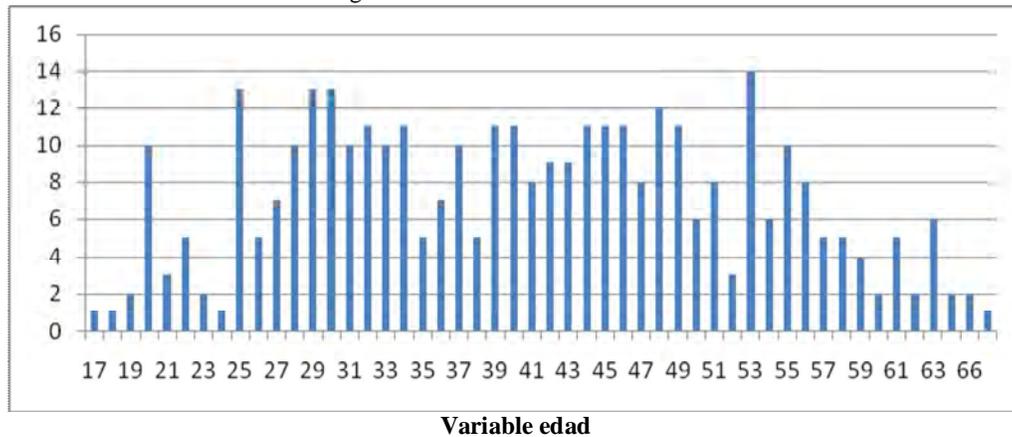
2.4. Depuración de registros:

Con respecto a la limpieza de datos del proceso de descubrimiento del conocimiento se puede determinar que no es necesario aplicar técnicas en relación a los datos *missing*, dado que la tabla tiene datos válidos en cada uno de sus campos.

Por otro lado, se realizará el análisis estadístico de cada variable para identificar su comportamiento individual y tratar de encontrar *outliers* o datos erróneos que puedan sesgar los resultados.

- **Edad:** Este atributo es cuantitativo o numérico y representa la edad del trabajador accidentado.

Figura 2.1. Distribución de la variable Edad



Media: 41.

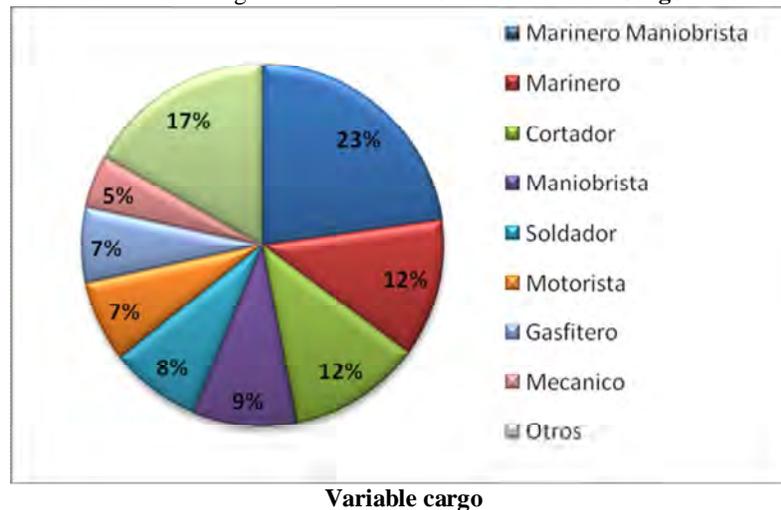
Moda: 53, 25, 29, 30.

Desviación estándar: 11.77

En este atributo se puede observar una distribución con cierta tendencia de frecuencia de casos en trabajadores cuyas edades oscilan entre los 25 y 55 años. Los valores más altos están en los trabajadores cuyas edades se ubican entre los 29 y 31 años, también se aprecian esta tendencia en edades entre 41 y 46 años. Por otro lado, hay valores atípicos en las edades de 25 y 53 años pues también están dentro de los valores de frecuencia más altos. En general la distribución parece seguir una tendencia oscilante y los valores atípicos no son muchos ni tienen gran impacto como para justificar su modificación.

- Cargo: Este atributo es cualitativo o categórico y representa el cargo según el organigrama de IMI, del trabajador accidentado.

Figura 2.2. Distribución de la variable Cargo

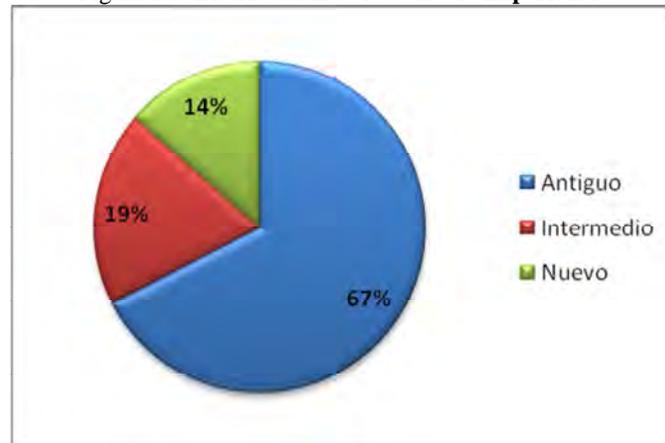


Se observa claramente una tendencia superior del cargo Marinero Maniobrista el cual encabeza las posiciones con un 23% de ocurrencias. Posteriormente se observa que las frecuencias se distribuyen a partir del 12%. Un dato interesante con respecto a esta información es referida a que todos los cargos (salvo el rubro Otros) tienen relación con la manipulación de equipos y/o herramientas industriales. Considérese que no toda la población laboral de IMI tiene labores manuales o mecánicas. A saber, poco menos del 80% de la actividad laboral tiene algún tipo de relación con equipos y/o herramientas.

Por otro lado, el Marinero Maniobrista es un puesto ubicado en zonas de alto riesgo, básicamente en las barcasas de trabajo de IMI (artefactos navales no propulsados que sirven de soporte para realización de trabajos industriales en plataformas petroleras marinas). También puede desarrollar actividades en plataformas petroleras.

- Experiencia: Este atributo es cualitativo o categórico y define el nivel de experiencia que tiene el trabajador en la labor que realiza. Existen 03 categorías: Personal nuevo cuyo periodo es de 03 a 06 meses; personal intermedio, cuyo periodo es de 06 a 12 meses; y finalmente, personal antiguo cuyo periodo es a partir de 01 año de trabajo continuo en un cargo en particular. Esta categorización ha sido determinada por IMI.

Figura 2.3. Distribución de la variable Experiencia

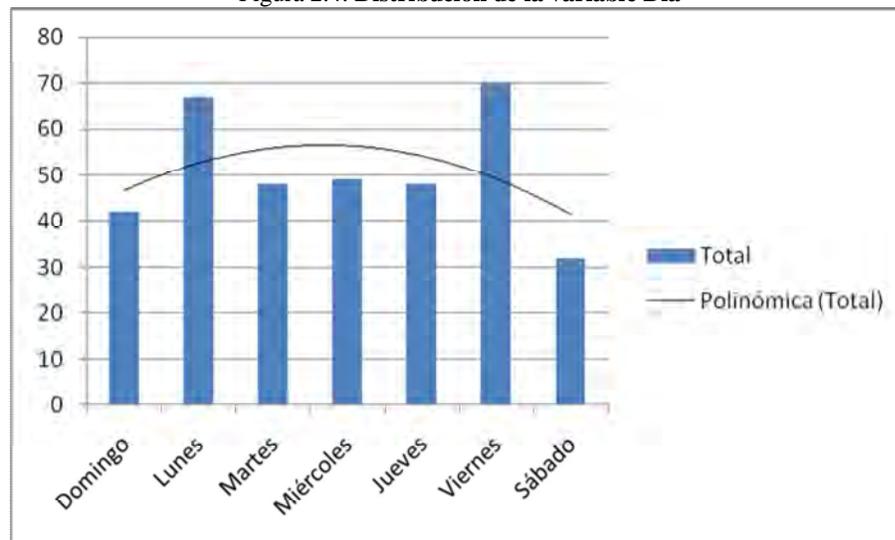


Variable experiencia

Se puede apreciar una alta tasa de incidencia en el personal antiguo, esto es en aquel personal que se supone está familiarizado con las labores que desempeña. En tasas menores se ubica al personal intermedio y nuevo.

- Día: Este atributo es cualitativo o categórico y representa el día en la cual ocurrió el accidente.

Figura 2.4. Distribución de la variable Día

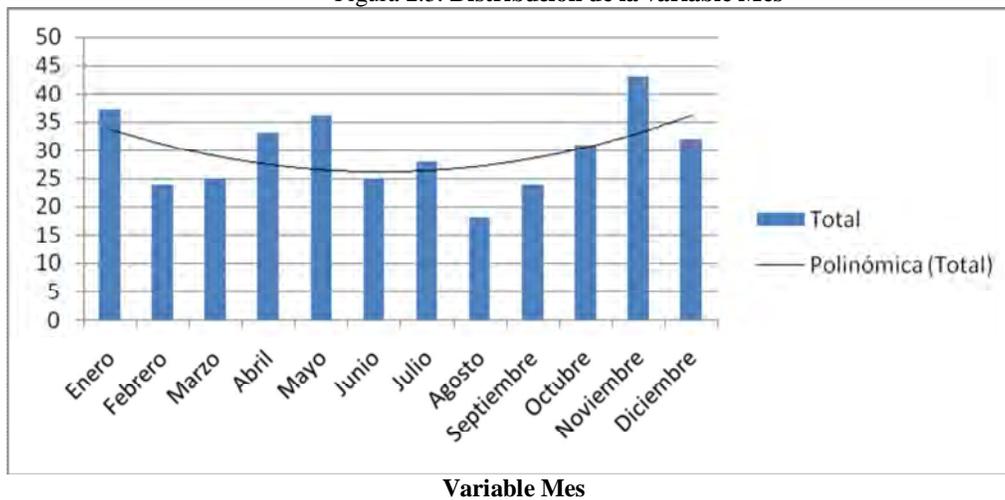


Variable día

En los datos se observa una distribución muy peculiar pues la mayor tasa de accidentabilidad se concentra en los días laborables; sin embargo y en general la fuerza laboral de IMI trabaja los 365 días del año. En todo caso, se podría considerar un comportamiento estacional en la que se observa mayor incidencia entre los días lunes y viernes; y menor, para los días de fin de semana. Se ha realizado un ajuste a través de una distribución de tendencia polinómica.

- Mes: Este atributo es cualitativo o categórico y representa el mes en la cual ocurrió el accidente.

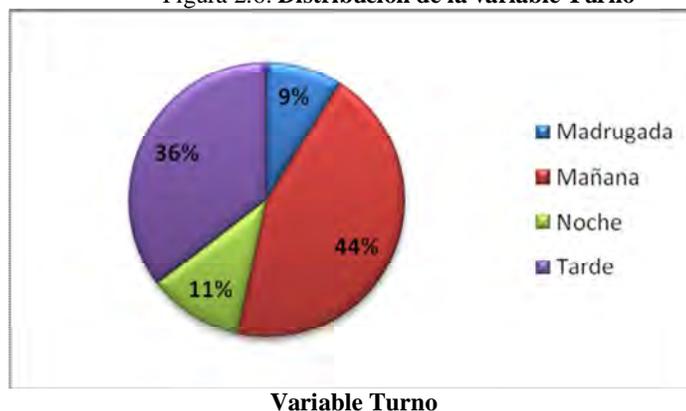
Figura 2.5. Distribución de la variable Mes



Se puede observar un comportamiento aparentemente aleatorio; sin embargo, se observa una tendencia posiblemente esperable en la que ocurren accidentes, luego hay una disminución para luego volver a registrar un incremento y así sucesivamente hasta finalizar el año. En general y con ayuda de un ajuste de distribución polinómico se puede apreciar que en los meses centrales del año hay una disminución de la accidentabilidad mientras que en el inicio y término de mismo, la tendencia es a aumentar.

- Turno: Este atributo es cualitativo o categórico y representa una subdivisión a los dos tipos de turno o jornada laboral (diurno y nocturno) en la cual pudo haber ocurrido el accidente. La distribución según IMI es la siguiente: turno diurno (mañana) a partir de 06:01 a.m. a 12:00 p.m.; turno diurno (tarde) a partir de las 12:01 p.m. a 06:00 p.m.; turno nocturno (noche) a partir de las 06:01 p.m. a 12:00 p.m. y turno nocturno (madrugada) a partir de las 12:01 p.m. a 06:00 a.m.
Esta subdivisión de los turnos de trabajo nace con la finalidad de facilitar la clasificación del horario en la que ocurre el accidente.

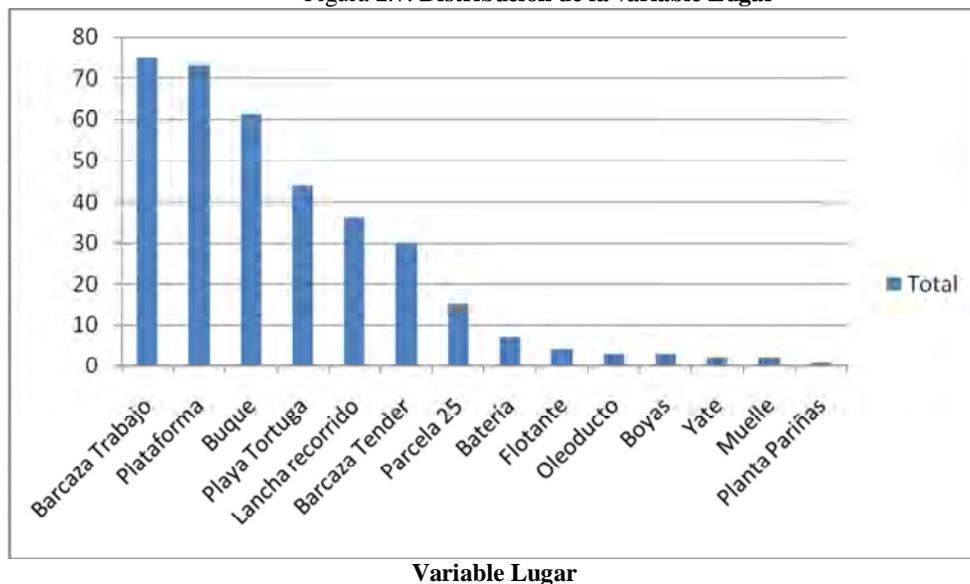
Figura 2.6. Distribución de la variable Turno



Se puede apreciar dos tendencias claras de las ocurrencias, mañana y tarde (ambos del turno diurno), mientras que en el turno nocturno la incidencia disminuye considerablemente. Ciertamente la fuerza laboral de IMI se desempeña mayoritariamente en el turno diurno.

- **Lugar:** Este atributo es cualitativo o categórico representa el lugar donde ocurrió el hecho. Cabe resaltar que existen varios atributos de la tabla de incidentes de IMI que detallan el lugar exacto donde ocurrió el accidente; sin embargo el grado de detalle hace que la descripción sea más compleja y que la cantidad de instancias de los atributos sea mayor. Por esta razón se escogió únicamente el atributo Lugar pues representa las categorías de los lugares donde han ocurrido los accidentes, tales como: barcasas, barcos, muelles, etc.

Figura 2.7. Distribución de la variable Lugar

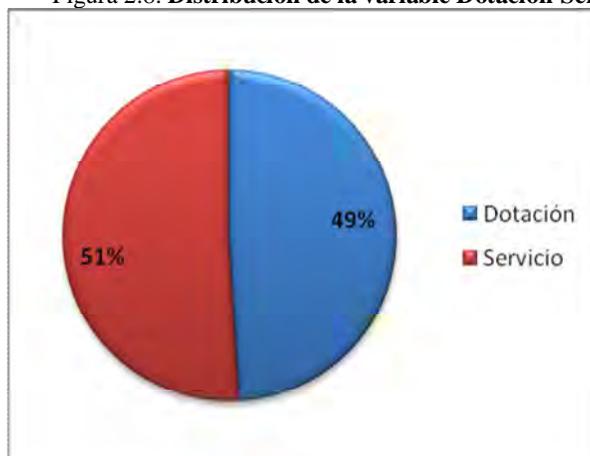


En la distribución se puede apreciar una clara tendencia de hechos que ocurren en barcazas de trabajo, plataformas petroleras y buques. Es interesante observar la alta incidencia de hechos que ocurren en la operación marina pues apenas se registran 05 categorías de sucesos en tierra: Parcela 25 (es un gran taller que contiene un muelle de trabajo), batería (locación donde se almacena el crudo), oleoducto (lugar por donde circula el petróleo crudo a través de líneas o ductos), muelle (lugar donde atracan los artefactos navales) y planta Pariñas (nombre de la empresa que procesa gas).

- **Tierra-Mar:** Este atributo es cualitativo o categórico y describe si el accidente ocurrió en el mar o tierra. En el mar se encuentran los artefactos navales llámese buques, barcos, naves y/o barcazas, también se tienen plataformas petroleras. Tierra se refiere a los distintos talleres ubicados en las instalaciones de la empresa en tierra. Normalmente el personal de estos talleres realiza trabajos en barcos acoderados a muelle o trabajos de mecánica dentro de los talleres adecuados para estos trabajos.
- **Dotación-Servicio:** Este atributo es cualitativo o categórico y es útil para diferenciar si el trabajador que resultó accidentado correspondía al grupo de trabajo del lugar donde se realizaba sus labores, o se trataba de un tercero que llegó proveniente de otro lugar

de trabajo. Esto ocurre a menudo en las embarcaciones en donde personal que viene de los talleres de soporte en tierra sube a bordo de la nave a hacer trabajos específicos (eléctricos, carpintería, etc.).

Figura 2.8. Distribución de la variable Dotación-Servicio

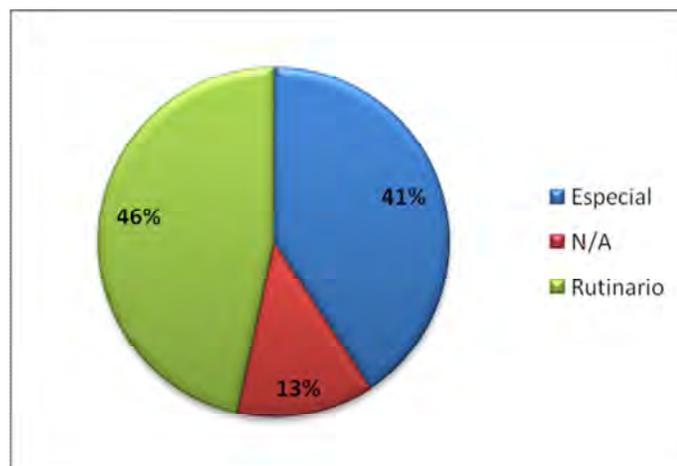


Variable Dotación-Servicio

En la distribución se puede apreciar que la ocurrencia de accidentes entre estas dos grandes categorías de personal es prácticamente la misma.

- Tipo de trabajo: Este atributo es cualitativo o categórico y diferencia el tipo de trabajo a realizar pues muchos trabajos son rutinarios o que siempre se realizan, mientras que algunos son específicos o particulares y su frecuencia es menor. Por otro lado, existe un tercer estado llamado “No Aplica” o NA que distingue aquellos casos en las cuales ocurre el accidente en circunstancias en que el trabajador no realizaba ningún trabajo. Vale decir que según la ley peruana en seguridad, referida en capítulos anteriores, menciona que un accidente de trabajo también es aquel que ocurra en la locación de trabajo aunque no se realice trabajo alguno.

Figura 2.9. Distribución de la variable Tipo de trabajo



Variable Tipo de Trabajo

Aunque hay una superioridad de casos en las cuales se hacían trabajos rutinarios no es muy marcada con respecto a trabajos especiales o no muy frecuentes. Por otro lado hay un margen no tan pequeño con respecto a casos en las cuales el trabajador se lesiona sin haber realizado labor alguna. Esto tiene sentido debido a que el trabajador está expuesto a los riesgos inherentes a su locación de trabajo los cuales se manifiestan en cualquier momento si es que no han sido controlados. Por ejemplo, un trabajador puede aprisionarse un dedo con la puerta del taller en el que trabaja.

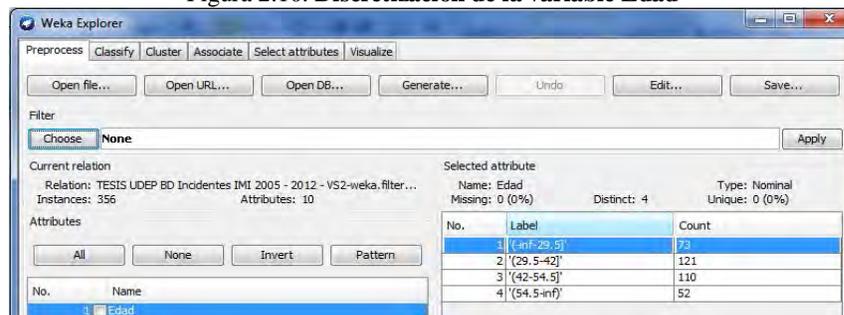
2.5. Limpieza y transformación

Luego de haber identificado los atributos de interés y haber analizado estadísticamente su comportamiento individual es necesario aplicar técnicas de limpieza y transformación a fin de obtener el *dataset* definitivo sobre el cual se va a trabajar.

Con respecto a la limpieza de datos se ha podido apreciar que en la distribución de datos de cada atributo no se observan valores atípicos que generen gran impacto sobre la media global de cada uno de ellos. Por otro lado, en lo que respecta a la transformación de datos en cada atributo se aprecia que se pueden realizar algunas optimizaciones.

- Edad: Para este atributo conviene realizar la técnica de discretización con el objeto de agrupar los valores en intervalos previamente determinados. Se utilizara el filtro *discretize* del *Weka* para desarrollar esta tarea.

Figura 2.10. Discretización de la variable Edad



Filtro *discretize*

Para este caso se ha definido el valor de *bins* (cubos) o intervalos de cuatro (04), los valores de $-\infty$ y $+\infty$ serán considerados según las edades de ley para laborar esto es: 17 años (considerando a los practicantes) y 70 años (edad de jubilación), respectivamente.

- Cargo: No se realizarán técnicas de transformación y limpieza pues no es necesario.
- Experiencia: No se realizarán técnicas de transformación y limpieza pues no es necesario.
- Día: Las variables temporales son cíclicas y ello puede generar problemas al momento de realizar el agrupamiento ya que el primer y último valor será representados como lejanos entre sí. Por ejemplo, el día lunes se encuentra cerca al martes pero lejos del sábado. La idea es que minimizar dicho impacto. Para resolver este problema se ha decidido empezar la serie por el día que tiene el menor valor quedando la distribución de la siguiente forma:

Tabla 2.2. Nuevo estado de la variable Día

Día	Nuevo estado
Sábado	1
Domingo	2
Lunes	3
Martes	4
Miércoles	5
Jueves	6
Viernes	7

Fuente: El Autor

Nótese que el cambio de estado implica el cambio de tipo de dato de la variable, pasando ésta a ser numérica.

- Mes: Este atributo presenta el mismo comportamiento que la anterior variable de modo que se realizará la misma tarea de transformación. En ese sentido, la serie iniciará con el mes que contiene el menor valor.

Tabla 2.3. Nuevo estado de la variable Mes

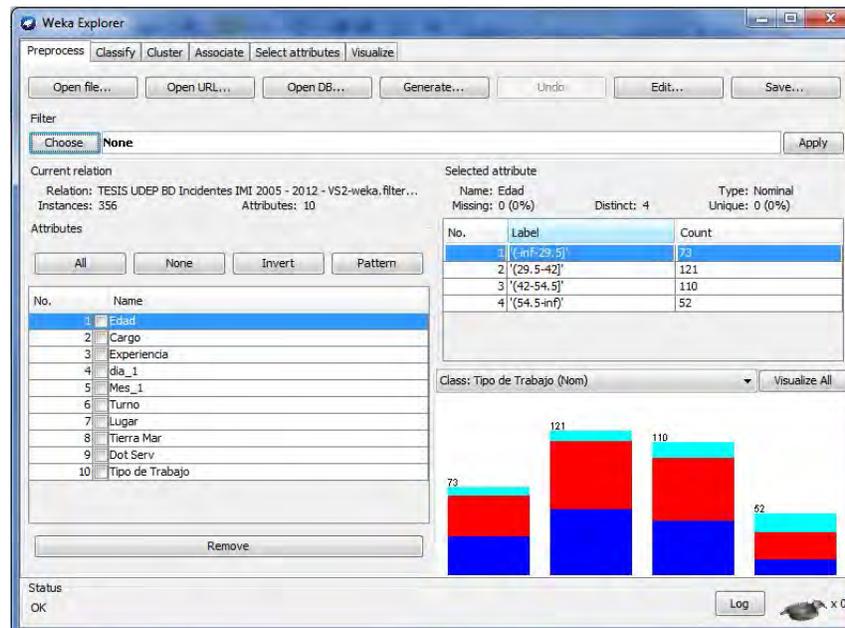
Día	Nuevo estado
Agosto	1
Septiembre	2
Octubre	3
Noviembre	4
Diciembre	5
Enero	6
Febrero	7
Marzo	8
Abril	9
Mayo	10
Junio	11
Julio	12

Fuente: El Autor

- Turno: No se realizarán técnicas de transformación y limpieza pues no es necesario.
- Lugar: No se realizarán técnicas de transformación y limpieza pues no es necesario.
- Dotación-Servicio: No se realizarán técnicas de transformación y limpieza pues no es necesario.
- Tipo de trabajo: No se realizarán técnicas de transformación y limpieza pues no es necesario.

2.6. Dataset final:

El *dataset* a trabajar en el *Weka* estará compuesto por 10 atributos y 356 registros. Cargados en el *Weka* quedarán de la siguiente forma:

Figura 2.11. *Dataset cargado en el Weka**Dataset cargado en Weka*

3. Desarrollo del experimento

El experimento a realizar estará conformado por los siguientes pasos:

1. Aplicar el algoritmo *K-means* en la tarea de *clustering* (agrupamiento) para encontrar patrones comunes en las variables de estudio.
2. Interpretar los resultados obtenidos haciendo uso de las herramientas que ofrece el *Weka* y bajo la opinión de expertos.
3. Aplicar el algoritmo de inducción C4.5 en la tarea de clasificación para identificar reglas de pertenencias de un valor a cada grupo.
4. Obtener conclusiones finales.

3.1. *Clustering* (agrupamiento):

Utilizando el algoritmo *K-means* el cual se encuentra disponible dentro del set de herramientas del *Wekas* e procede a realizar esta tarea considerando 3 grupos y un máximo de 500 iteraciones. La función de distancia quedará por defecto en distancia euclídea.

Con respecto al valor de la semilla (valor inicial que requiere el algoritmo para iniciar sus propias iteraciones) se escogerá aquella que tenga menor error cuadrático luego de 10 iteraciones. Este procedimiento es heurístico. Se resaltarán en amarillo la semilla escogida.

Tabla 3.1. Iteraciones para encontrar la semilla más adecuada con 3 cluster.

Semilla	Error cuadrático
1	1263.48
2	1235.15
3	1221.62
4	1278.58
5	1237.55
6	1281.70
7	1275.09
8	1224.61
9	1252.98
10	1247.20

Fuente: El autor.

Utilizando entonces, la semilla 3 se obtienen los siguientes resultados en *Weka*.

Figura 3.1. Resultado del algoritmo *K-means* en *Weka*

```

Cluster centroids:
Attribute          Full Data          Cluster#
                   (356)              0                  1                  2
                   (356)              (166)              (132)              (58)
-----
Edad                '(29.5-42]''       '(29.5-42]''       '(42-54.5]''       '(-inf-29.5]''
Cargo               Marinero Maniobrista Marinero Maniobrista Marinero Maniobrista Soldador
Experiencia         Antiguo             Antiguo             Antiguo             Intermedio
dia_1               4.3034             4.7349             3.7121             4.4138
Mes_1               6.5983             7.0482             6.0606             6.5345
Turno               Mañana             Tarde              Mañana             Mañana
Lugar               Barcaza Trabajo   Barcaza Trabajo   Plataforma         Playa Tortuga
Tierra Mar         Mar                Mar                Mar                Tierra
Dot Serv           Servicio           Dotación           Servicio           Servicio
Tipo de Trabajo    Rutinario         Rutinario         Especial            Especial

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances
0      166 ( 47%)
1      132 ( 37%)
2       58 ( 16%)

```

Clustering

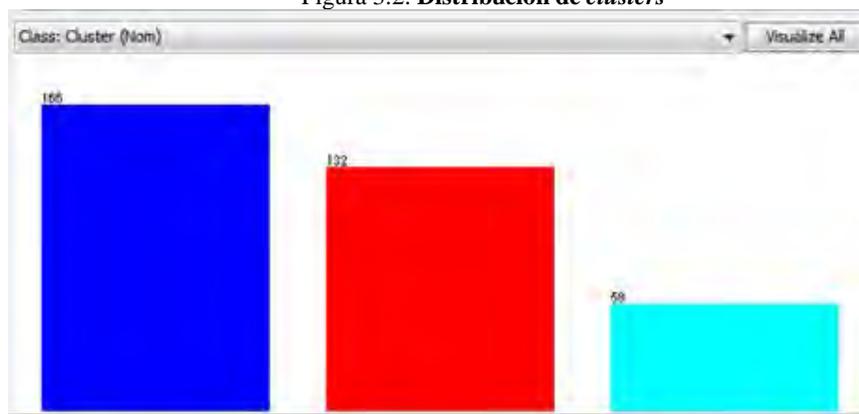
A partir de estos resultados se puede resumir en una tabla aquellos más representativos:

Tabla 3.2. Valores más representativos.

Atributo	Instancia
Edad	(29,5-42]
Cargo	Marinero Maniobrista
Experiencia	Antiguo
Día_1	4 (Martes)
Mes_1	7 (Febrero)
Turno	Mañana
Lugar	Barcaza Trabajo
Tierra Mar	Mar
Dot Serv	Servicio
Tipo de Trabajo	Rutinario

Fuente: El autor.

A continuación se procede a realizar la evaluación de cada variable con respecto a cada *cluster*.

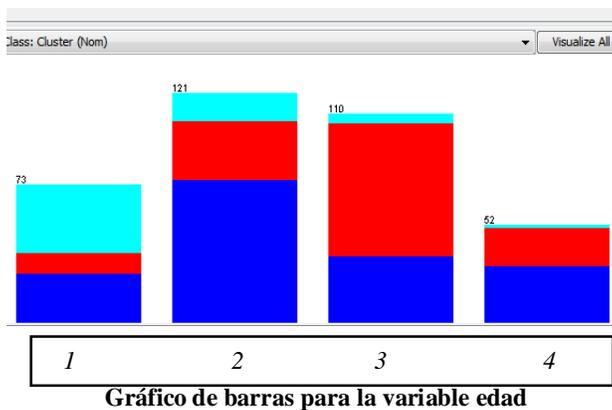
Figura 3.2. Distribución de *clusters*Gráfico de barras de los *clusters*

Las proporciones de los *clusters* son las siguientes: *cluster* 0 con 47%, *cluster* 1 con 37% y *cluster* 2 con 16%.

- Edad: Se aprecia que el intervalo 1 y 3 presentan un comportamiento distinto.

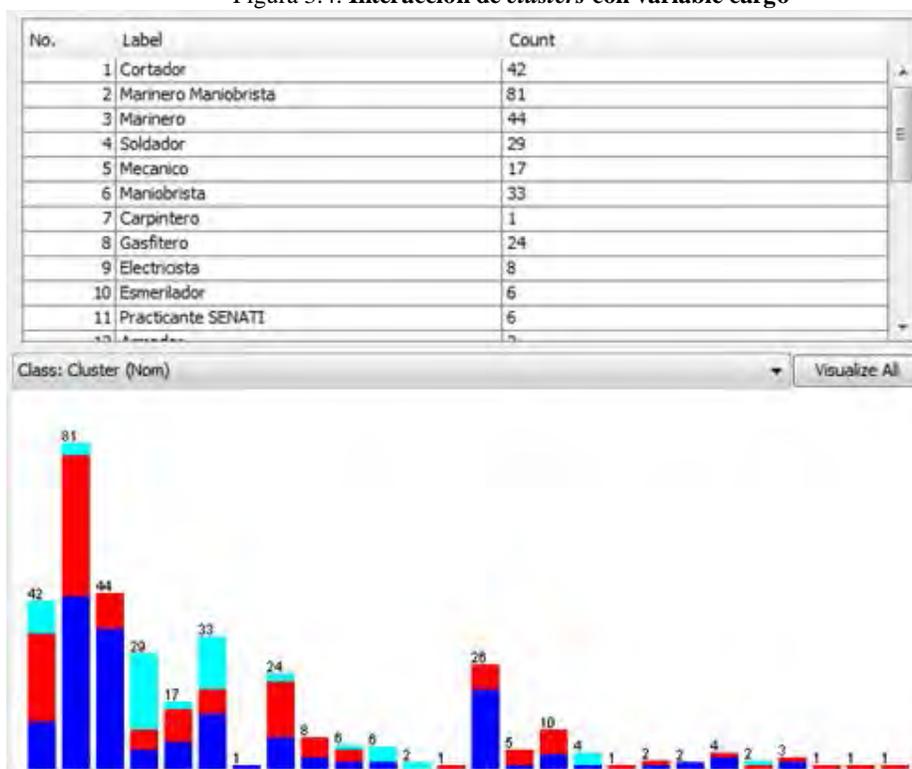
Figura 3.3. Interacción de *clusters* con variable edad

No.	Label	Count
1	'(-inf-29.5]'	73
2	'(29.5-42]'	121
3	'(42-54.5]'	110
4	'(54.5-inf)'	52



- Cargo: Varias de las instancias se asemejan a la distribución de los *cluster*.

Figura 3.4. Interacción de *clusters* con variable cargo



- Experiencia: Las distribuciones de las instancias no son tan parecidas a la de los *clusters*.

Figura 3.5. Interacción de *clusters* con variable experiencia

No.	Label	Count
1	Antiguo	240
2	Intermedio	68
3	Nuevo	48

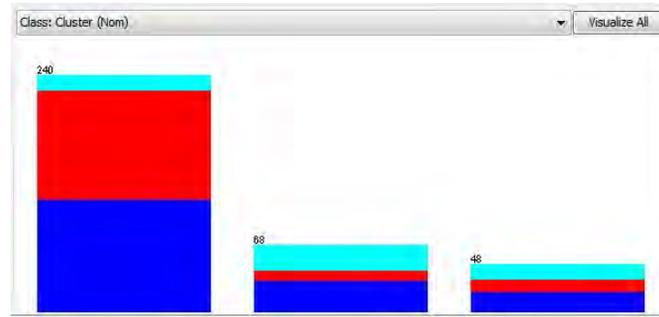


Gráfico de barras para la variable experiencia

- Día: Las distribuciones de las instancias son muy similares a la de los *clusters*. La distribución de estados es según tabla 2.4.

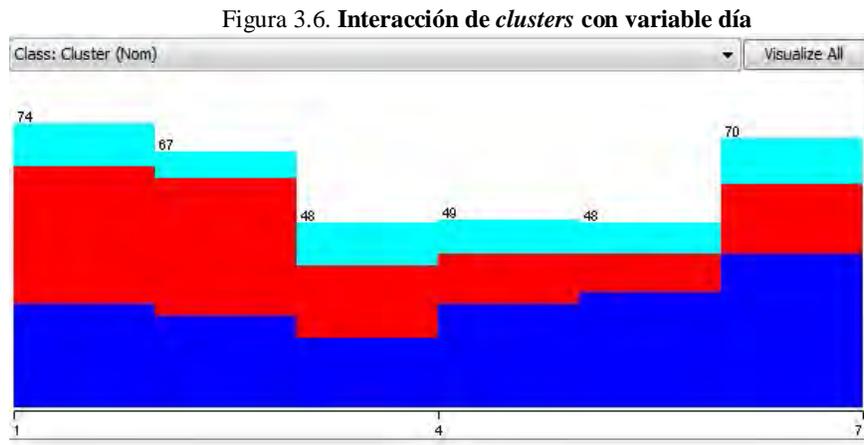


Gráfico de barras para la variable día

- Mes: Las distribuciones de las instancias son muy similares a la de los *clusters*. La distribución de estados es según tabla 2.5.

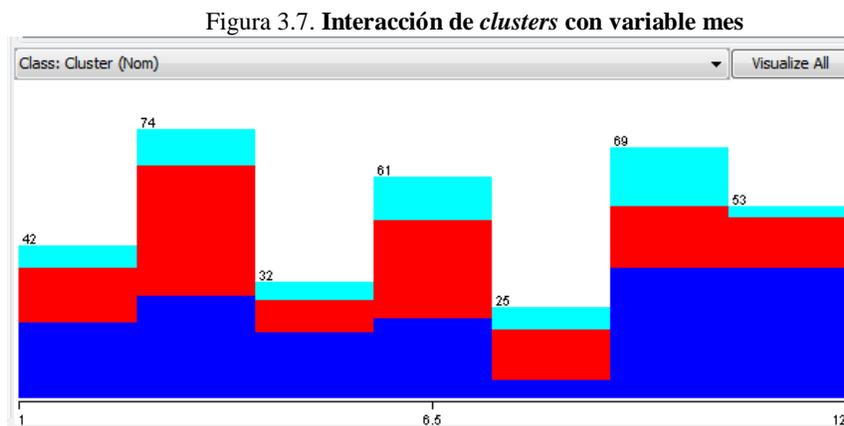


Gráfico de barras para la variable mes

- Turno: Salvo la primera instancia, las distribuciones son muy similares a la de los *clusters*.

Figura 3.8. Interacción de *clusters* con variable turno

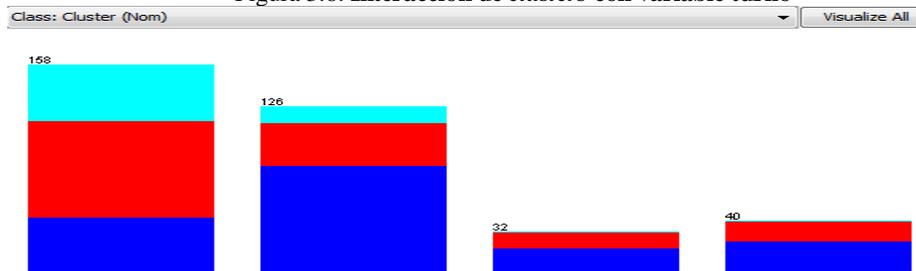


Gráfico de barras para la variable turno

- Lugar: Se aprecian distribuciones distintas a la de los *clusters*.

Figura 3.9. Interacción de *clusters* con variable lugar

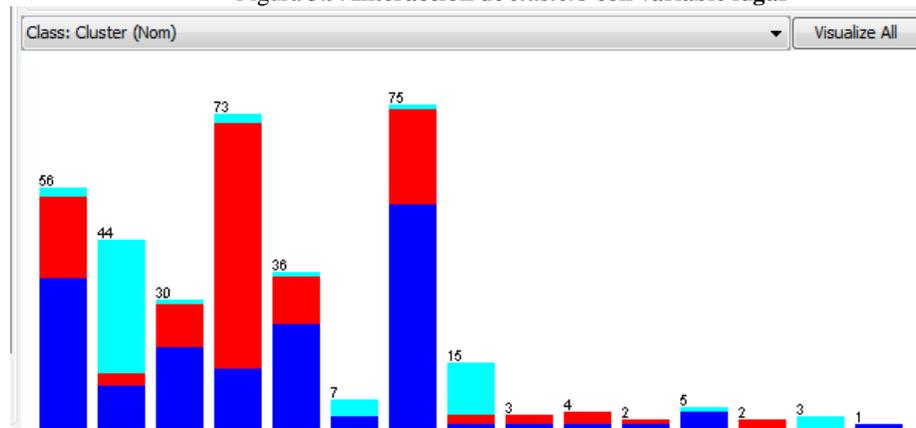


Gráfico de barras para la variable lugar

- Tierra Mar: La distribución no es muy proporcional.

Figura 3.10. Interacción de *clusters* con variable tierra mar

No.	Label	Count
1	Mar	290
2	Tierra	66

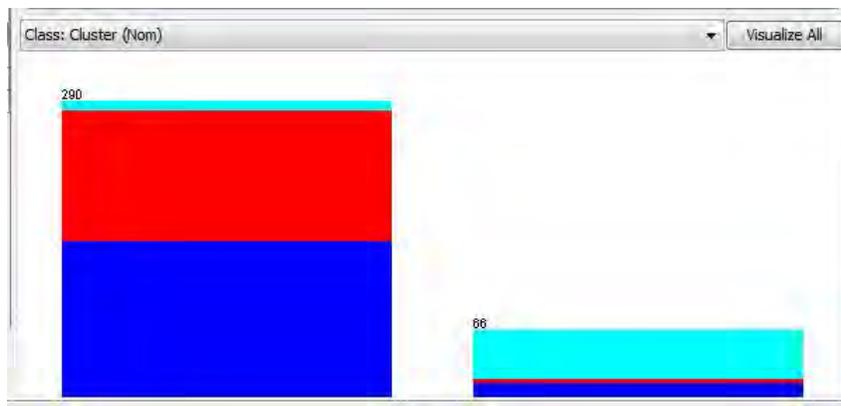


Gráfico de barras para la variable tierra mar

- Dotación Servicio: La distribución no es muy proporcional.

Figura 3.11. Interacción de *clusters* con variable dotación servicio

No.	Label	Count
1	Servicio	181
2	Dotación	174
3	Especial	1

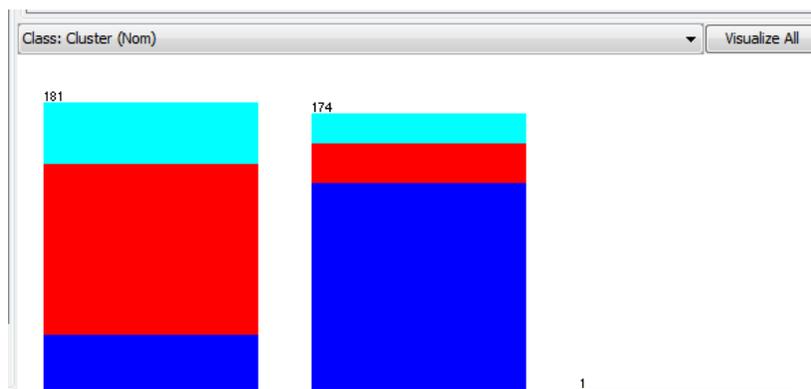


Gráfico de barras para la variable dotación servicio

- Tipo de trabajo: En las 3 instancias se observan comportamientos diferentes.

Figura 3.12. Interacción de *clusters* con variable tipo de trabajo

No.	Label	Count
1	Especial	145
2	Rutinario	166
3	N/A	45

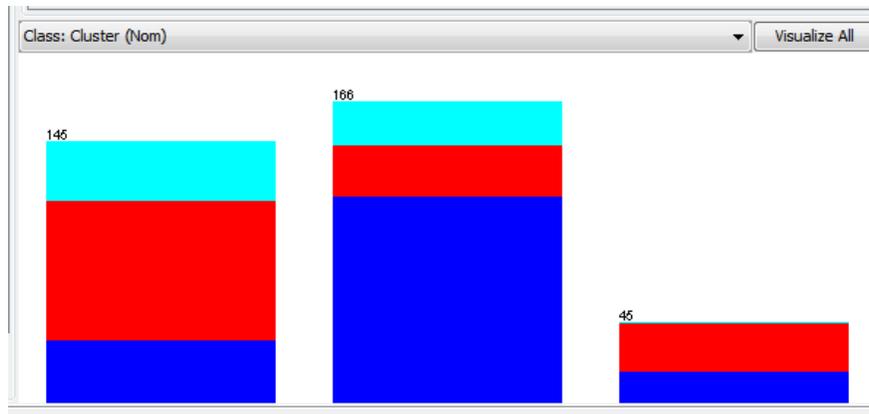


Gráfico de barras para la variable tipo de trabajo

Luego de realizar esta evaluación personalizada por cada atributo se debe evaluar aquellos en cuyas instancias se encuentran distribuciones de los *clusters* distintas al patrón individual de cada uno de ellos. Como se mencionó anteriormente, esto indica comportamientos especiales que deberían ser analizados en detalle. Así se tiene que los siguientes atributos tienen mayor interacción con los *clusters*: edad, cargo, turno, lugar, dot serv y tipo de trabajo.

Se utilizará la pestaña *visualize* del *Weka* que muestra gráficos de los atributos que se quieren comparar. En este caso se analizará la interacción de cada *cluster* con las variables más significativas según el análisis anterior. Se verificará la densidad de distribución de cada instancia de un atributo con respecto a cada *cluster*. Así se verá si las distribuciones de cada instancia con respecto a cada *cluster* son homogéneas o irregulares.

- Edad: Se puede observar que todos los *cluster* registran instancias más densas que otras. El *cluster* 2 tiene mayor densidad en la población de personal que tiene menos de 30 años. Su distribución es irregular. El *cluster* 1 tiene mayor densidad en la población de trabajadores cuyas edades están entre 42 y 55 años, se podría decir que su distribución es uniforme. Finalmente el *cluster* 0 es el más estable teniendo mayor densidad en trabajadores con edades entre 30 y 42 años. Recuérdese que la mayor tasa de accidentes se registra en edades entre 30 y 55 años.

Figura 3.13. Interacción de *clusters* con instancias de la variable edad

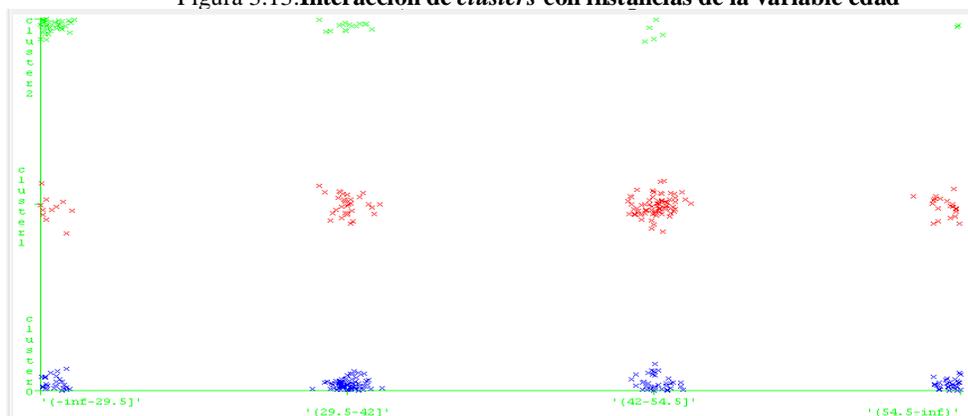


Gráfico de dispersión para la variable edad

- Cargo: Los 3 *clusters* registran mayor densidad en los cargos marintero maniobrista y cortador.

Figura 3.14. Interacción de *clusters* con instancias de la variable cargo

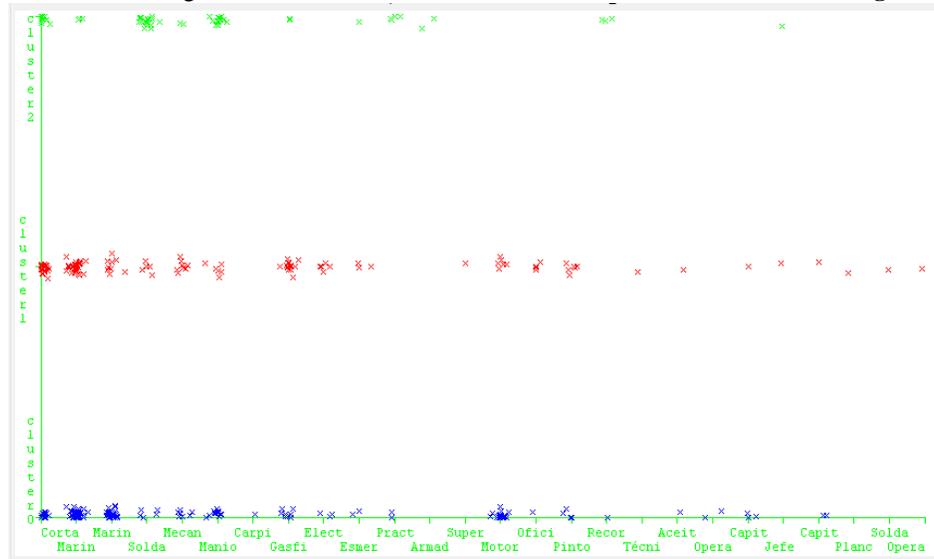


Gráfico de dispersión para la variable cargo

- Turno: Los 3 *clusters* registran mayor densidad en las instancias de mañana y tarde. La distribución disminuye aunque con una distribución algo uniforme para los casos de trabajos de noche y madrugada.

Figura 3.15. Interacción de *clusters* con instancias de la variable turno

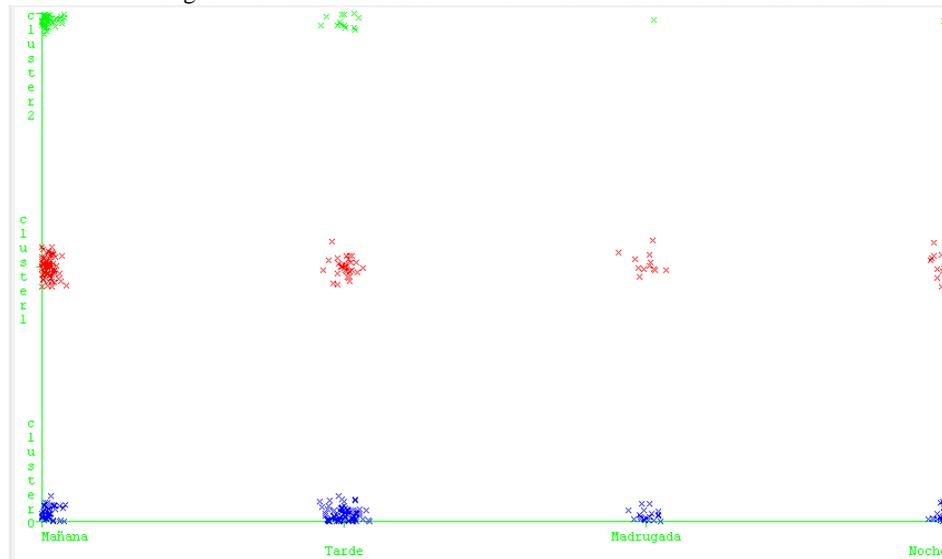


Gráfico de dispersión para la variable turno

- Lugar: Los 3 *clusters* registran mayor densidad en distintas instancias. Para el *cluster* 2 la mayor incidencia se registra en Playa Tortuga. Para el *cluster* 1 la densidad es mayor en Plataformas y finalmente, en el *cluster* 0 la densidad mayor es en Barcazas. Interesante

identificar que los *clusters* 1 y 0 registran mayores densidades en trabajos realizados en el mar.

Figura 3.16. Interacción de *clusters* con instancias de la variable lugar

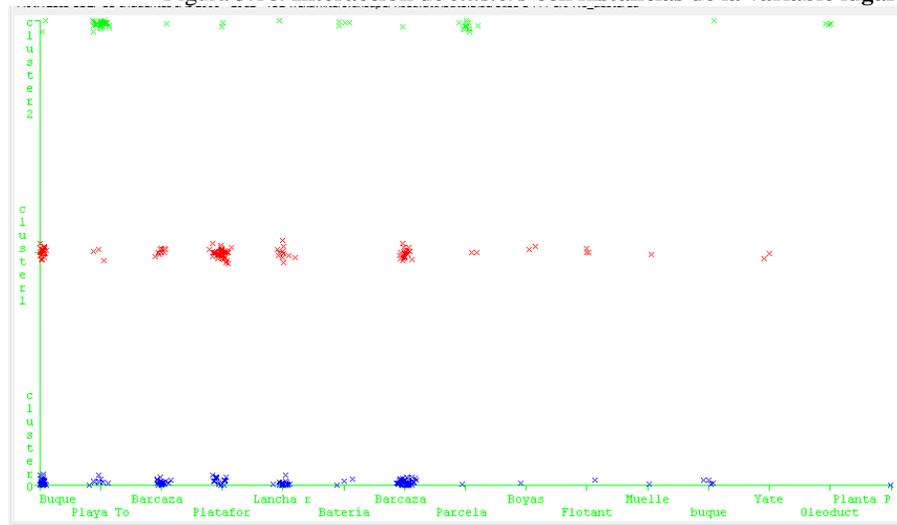


Gráfico de dispersión para la variable lugar

- Dotación Servicio: Tanto el *cluster* 2 y 1 registran altas densidades en la instancia que corresponde a trabajadores de servicio; es decir aquellos que asisten a una locación remota a realizar trabajos solicitados.

Figura 3.17. Interacción de *clusters* con instancias de la variable dotación servicio

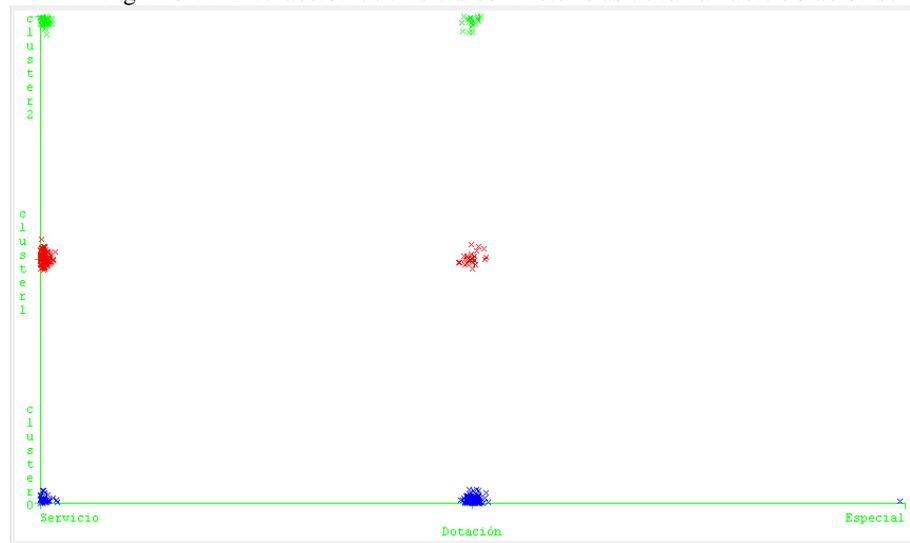


Gráfico de dispersión para la variable dotación servicio

- Tipo de trabajo: Tanto para los *cluster* 2 y 1 los trabajos especiales tienen mayor densidad mientras que para el *cluster* 0 la mayor densidad se registra en trabajos rutinarios. Las distribuciones son uniformes en los 3 casos.

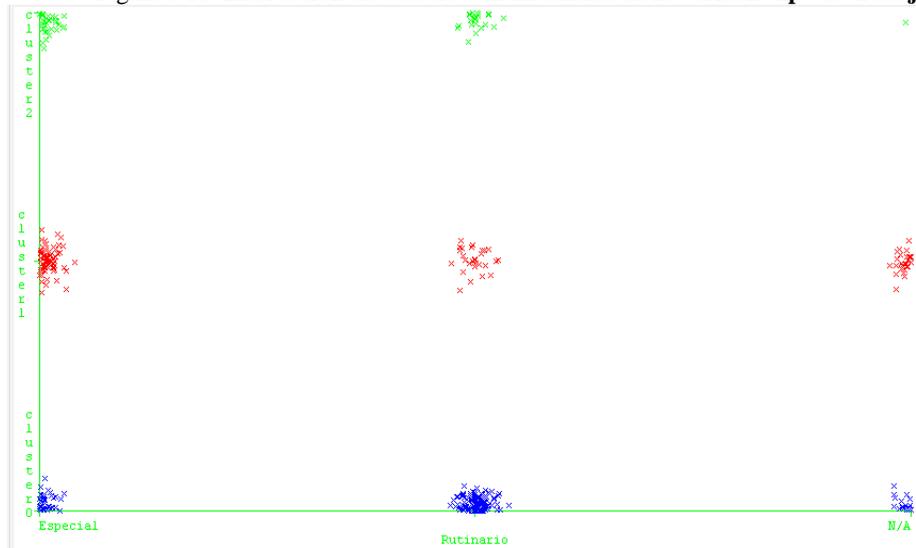
Figura 3.18. Interacción de *clusters* con instancias de la variable tipo de trabajo

Gráfico de dispersión para la variable tipo de trabajo

Ahora se podría utilizar esta herramienta para revisar las interrelaciones entre atributos en los cuales se tiene particular interés. En ese sentido, convendría evaluar la edad con la variable tierra mar a fin de evaluar si la accidentabilidad en mar o tierra podría estar condicionada por la edad del trabajador; así mismo el lugar con el turno pues se sabe que es en las locaciones de mar el personal pernocta teniendo mayores posibilidades de un accidente de trabajo.

- Interrelación edad – tierra mar: Hay una notable interacción entre los *cluster* 1 y 0 para los casos de accidentes en edades entre 30 y 55 años ocurridos en el mar. Parece ser que esa es la tendencia que caracteriza a ambos *cluster*. Por otro lado el *cluster* 2 marca una clara ocurrencia de casos en tierra y particularmente mayor para personal menor de 30 años.

Figura 3.19. Interrelación de variables edad y tierra mar

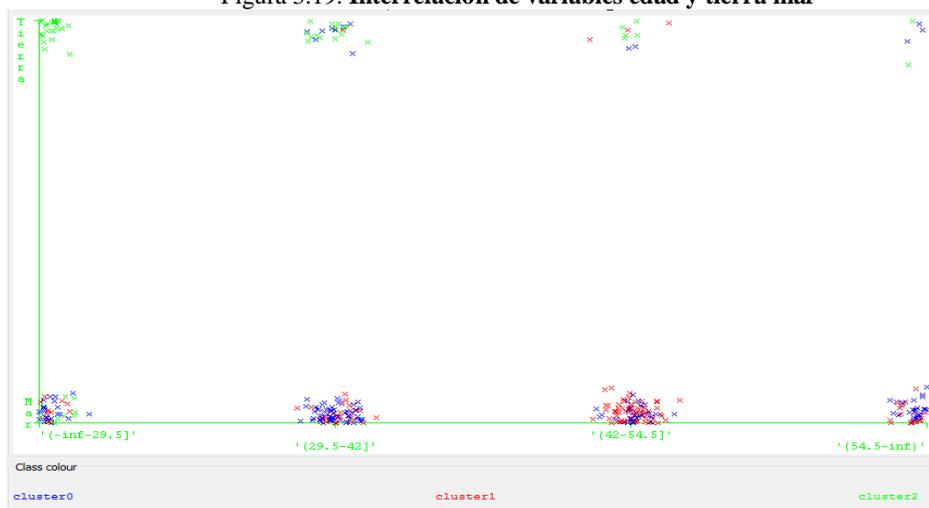


Gráfico de dispersión para evaluar la interacción de las variables tierra mar y edad

- Interrelación lugar – turno: Se puede apreciar una buena interacción de los *cluster* 0 y 1 para los casos de accidentes registrados en barcazas y en la mañana así como en plataformas y ocurridos en la tarde.

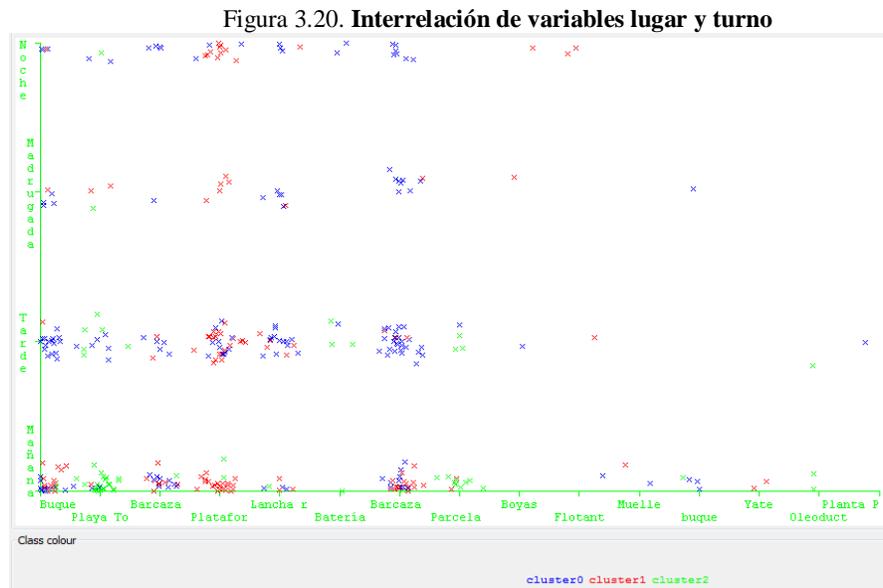


Gráfico de dispersión para evaluar la interacción de las variables turno y lugar

Primera interpretación.

En función de estos análisis se pueden abordar las primeras conclusiones con respecto a los *clusters* obtenidos. Así se tiene:

1. *Cluster* 0.- Caracterizado por accidentes de trabajo que ocurren en el mar, principalmente en barcazas a la tripulación de la misma. En general ocurren en turno diurno (mañana o tarde) y a personal que realiza los trabajos de maniobra (maniobristas).
2. *Cluster* 1.-Caracterizado por accidentes de trabajo que ocurren en el mar, principalmente en plataformas petroleras. También ocurren en turno diurno y el personal afectado es el que realiza trabajos especiales (reparación de estructuras en plataformas).
3. *Cluster* 2.-Caracterizado por accidentes de trabajo que ocurren en tierra, básicamente en los talleres que tiene la empresa. Igualmente ocurren en turno diurno y el personal afectado normalmente tiene edades menores a 30 años.

3.2. Implementación del algoritmo de inducción C4.5 para la tarea de clasificación de los *clusters*

En la tabla descrita a continuación se describen los atributos más “influyentes” según los distintos métodos de evaluación utilizados. Se resaltan en colores aquellos atributos que figuran en la mayoría de métodos en una relación de mayor a menor.

Tabla 3.3. Selección de atributos.

Algoritmo de evaluación	<i>Wrapper</i>	<i>Classifier</i>	<i>Consistency</i>	<i>CfsSubsetEval</i>	<i>ChiSquare</i>	<i>Gain Ratio</i>
Método de búsqueda	<i>Exhaustive</i>	<i>Exhaustive</i>	<i>Exhaustive</i>	<i>Exhaustive</i>	<i>Ranker</i>	<i>Ranker</i>
Edad	x	x	x	x	4	5
Cargo			x		3	
Experiencia	x	x				
Día		x	x	x		
Mes	x	x				
Turno			x	x		
Lugar			x	x	1	3
Tierra Mar	x	x		x	2	1
Dot-Serv	x	x	x	x		2
Tipo de trabajo	x	x	x	x	5	4

Fuente: El autor.

Se aprecia que todos los métodos coinciden en elegir los atributos edad y tipo de trabajo. Posteriormente se tiene los atributos tierra mar y dotación servicio. El atributo lugar sigue en la lista y finalmente los atributos cargo y día.

Dados los atributos preseleccionados se procede a correr el algoritmo C4.5 con cada combinación

Tabla 3.4. Aplicación del algoritmo C4.5.

Combinación	Atributos										Resultado
	Dot-Serv	Edad	Tipo de trabajo	Tierra Mar	Día	Lugar	Experiencia	Mes	Turno	Cargo	
1	X	X	X	X	X	X	X	X	X	X	92.4%
2	X	X	X	X	X	X	X	X	X		92.9%
3	X	X	X	X	X	X	X	X			93.2%
4	X	X	X	X	X	X	X				91.3%
5	X	X	X	X	X	X					87.6%

Fuente: El autor.

A continuación se procederá a revisar los resultados arrojados por *Weka* luego de aplicada la tarea de clasificación en la combinación óptima (combinación 03).

- El porcentaje de clasificación de las instancias es de 93.25% lo cual resulta bueno en términos generales de modo que los *clusters* obtenidos se corresponden con un patrón o criterio determinado en relación a sus datos.
- Según la matriz de confusión el *cluster 2* posee el valor más alto de asignación el cual lo hace más estable con respecto a los otros dos.

Figura 3.21. Resultados combinación 03

```

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      332          93.2584 %
Incorrectly Classified Instances    24           6.7416 %
Kappa statistic                    0.8909
Mean absolute error                 0.0748
Root mean squared error             0.1933
Relative absolute error             18.1172 %
Root relative squared error         42.578 %
Total Number of Instances          356

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.946   0.074   0.918     0.946   0.932     0.975   cluster0
                0.902   0.031   0.944     0.902   0.922     0.975   cluster1
                0.966   0.01    0.949     0.966   0.957     0.998   cluster2
Weighted Avg.   0.933   0.048   0.933     0.933   0.932     0.979

=== Confusion Matrix ===

 a  b  c  <-- classified as
157 7  2 | a = cluster0
 12 119 1 | b = cluster1
 2  0  56 | c = cluster2

```

Resultados de la tarea de clasificación (combinación 01)

3.3. Árbol de Clasificación

El árbol tiene 34 hojas (nótese que éste se encuentra “podado” siendo esta una función posible gracias al algoritmo C4.5) nace tomando como nodo principal al atributo “Tierra Mar” del cual se obtienen dos posibles instancias: tierra y mar. Para el caso de la instancia “Mar” el árbol deriva en el nodo “Dot Serv” el cual posee 03 instancias: servicio, dotación y especial, precisamente esta última concluye en una hoja, la misma que es el *cluster* 0. Interesante ver a este nivel, la profundidad del nodo “Experiencia” en su instancia “Antiguo” pues se definen 3 nodos más de manera sucesiva: “Tipo de Trabajo”, “Edad” y “Mes”. Dentro del nodo “Experiencia” en su instancia “Intermedio” se aprecia claramente cómo se definen todos los *clusters* en función del nodo “Edad” en todas las instancias de este último nodo, lo cual quiere decir que la ocurrencia de accidentes para este tipo de personal es indistinta de la edad del trabajador. Para el caso de la instancia “Dotación” que proviene del nodo “Dot Serv” se aprecia que figuran los nodos “Edad” y “Día” sucesivamente para definir el comportamiento de la accidentabilidad de este tipo de trabajadores (recuérdese que el personal de la dotación es aquel que pertenece directamente a la tripulación de la barcaza o buque en donde ocurre la incidencia).

Por otro lado en la instancia “Tierra” del nodo “Tierra Mar” el nodo subsiguiente es “Edad” en cuyo caso se aprecia que para trabajadores menores de 30 años (instancia [-inf-29.5]) corresponde el *cluster* 2 con una tasa de clasificación del 100%. Interesante ver además como hay mayor interacción de *clusters* y nodos para el caso de personal cuyas edades oscilen entre los 30 y 43 años.

A continuación se definen las reglas de clasificación mas significativas, es decir aquellas que agrupan una mayor cantidad de registros de cada *cluster* obtenidas en la aplicación del

algoritmo C4.5. En ese sentido se tiene que de todas las reglas obtenidas, 9 de ellas clasifican al 76% de instancias:

Regla 1

Si Tierra Mar=Mar
 Y Dot Serv= Servicio
 Y Experiencia=Antiguo
 Y Tipo de Trabajo=Especial
 ENTONCES *Cluster 1* (42.0)

Regla 2

Si Tierra Mar=Mar
 Y Dot Serv= Servicio
 Y Experiencia=Antiguo
 Y Tipo de Trabajo=Rutinario
 Y Edad= (42-54.5]
 ENTONCES *Cluster 1* (14.0/1.0)

Regla 3

Si Tierra Mar=Mar
 Y Dot Serv= Servicio
 Y Experiencia=Antiguo
 Y Tipo de Trabajo=N/A
 ENTONCES *Cluster 1* (24.0)

Nótese que en estas tres primeras reglas se define claramente el *cluster 1* para personal que trabaja en el mar, no pertenece a la dotación de la barcaza o barco, es antiguo y que realiza trabajos ya sean especiales o rutinarios. Hay un alto grado de clasificación para instancias que cumplan estas características.

Regla 4

Si Tierra Mar=Mar
 Y Dot Serv= Dotación
 Y Tipo de Trabajo=Especial
 Y Edad=(29.5-42]
 ENTONCES *Cluster 0* (18.0/1.0)

Regla 5

Si Tierra Mar=Mar
 Y Dot Serv= Dotación
 Y Tipo de Trabajo=Especial
 Y Edad=(42-54.5]
 ENTONCES *Cluster 1* (17.0/4.0)

En este par de reglas se observa que dada las mismas condiciones, personal que labora en mar en trabajos especiales se tienen dos posibilidades, si sus edades son entre 30 y 42 años corresponde el *cluster* 0 mientras que si estos tienen edades entre 42 y 55 años, corresponde el *cluster* 1. Recuerdese que mientras en el *cluster* 0 se registran casos de accidentes en barcazas o buques, en el *cluster* 1 se tienen accidentes en plataformas.

Regla 6

Si Tierra Mar=Mar
 Y Dot Serv= Dotación
 Y Tipo de Trabajo=Rutinario
 ENTONCES *Cluster* 0 (80.0/3.0)

Regla 7

Si Tierra Mar=Mar
 Y Dot Serv= Dotación
 Y Tipo de Trabajo=N/A
 ENTONCES *Cluster* 0 (19.0/2.0)

En estas reglas se tiene que en general para trabajadores que realizan labores en el mar en trabajos rutinarios o en accidentes que no tienen causa directa con un trabajo en particular (por ejemplo sufrir un accidente en la ducha) se registra con claridad el *cluster* 0.

Regla 8

Si Tierra Mar=Tierra
 Y Edad=(-inf-29.5]
 ENTONCES *Cluster* 2 (26.0)

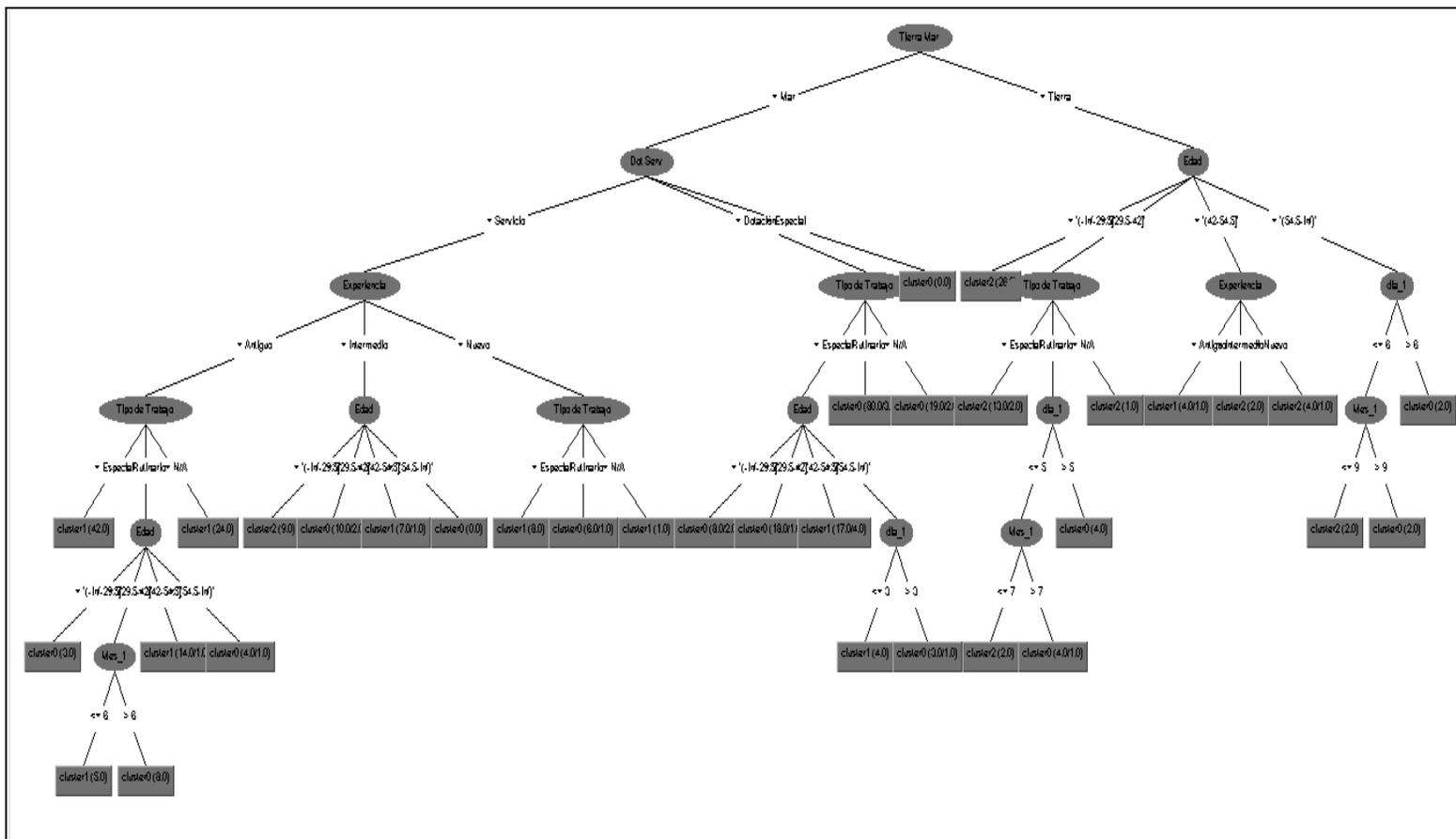
Regla 9

Si Tierra Mar=Tierra
 Y Tipo de Trabajo=Especial
 ENTONCES *Cluster* 2 (13.0/2.0)

Finalmente las reglas que definen el comportamiento de la accidentabilidad en tierra, caracterizado por personal con edades menores a 30 años y realizando trabajos especiales.

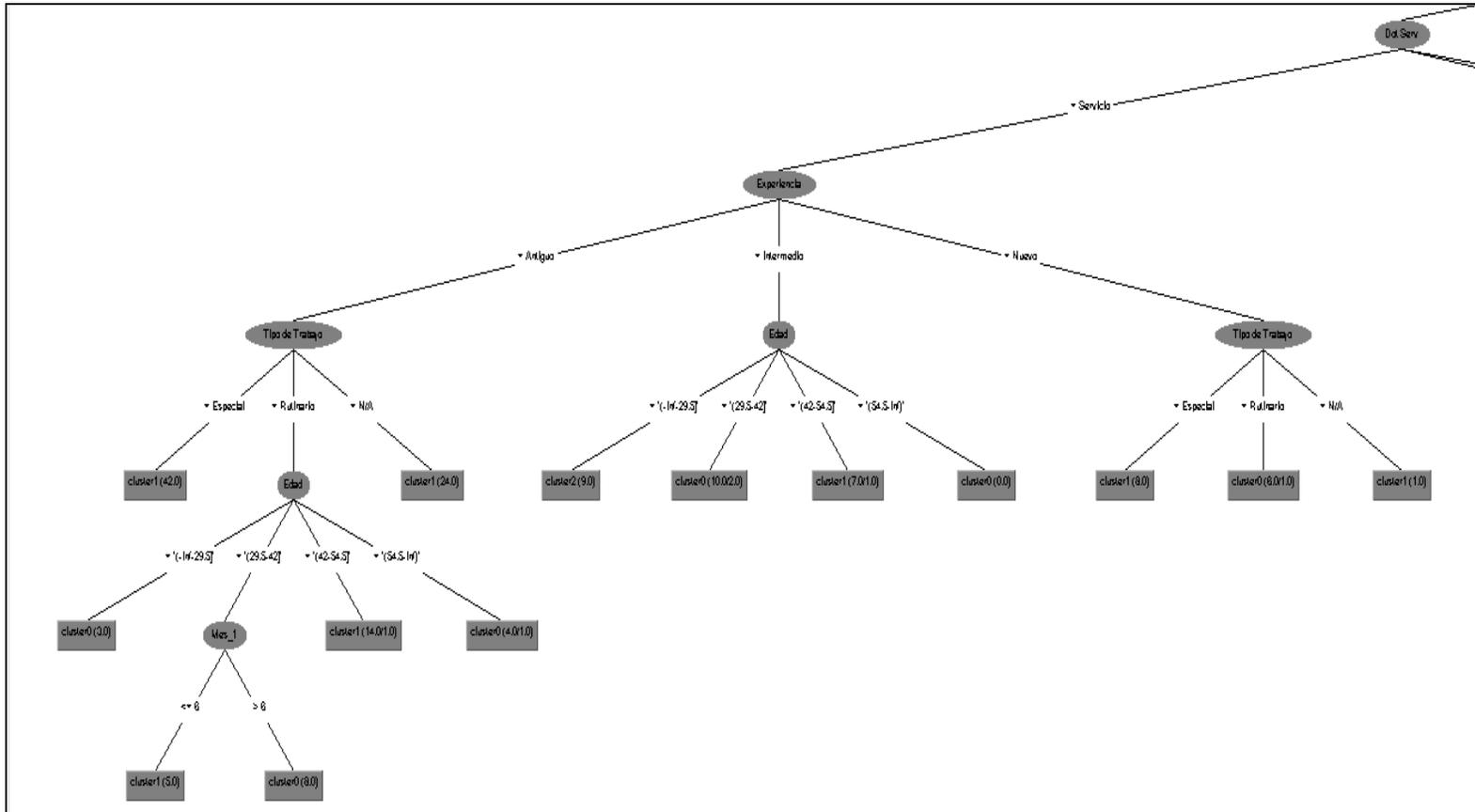
A continuación se muestran las gráficas del árbol.

Figura 3.22. Resultados combinación 03



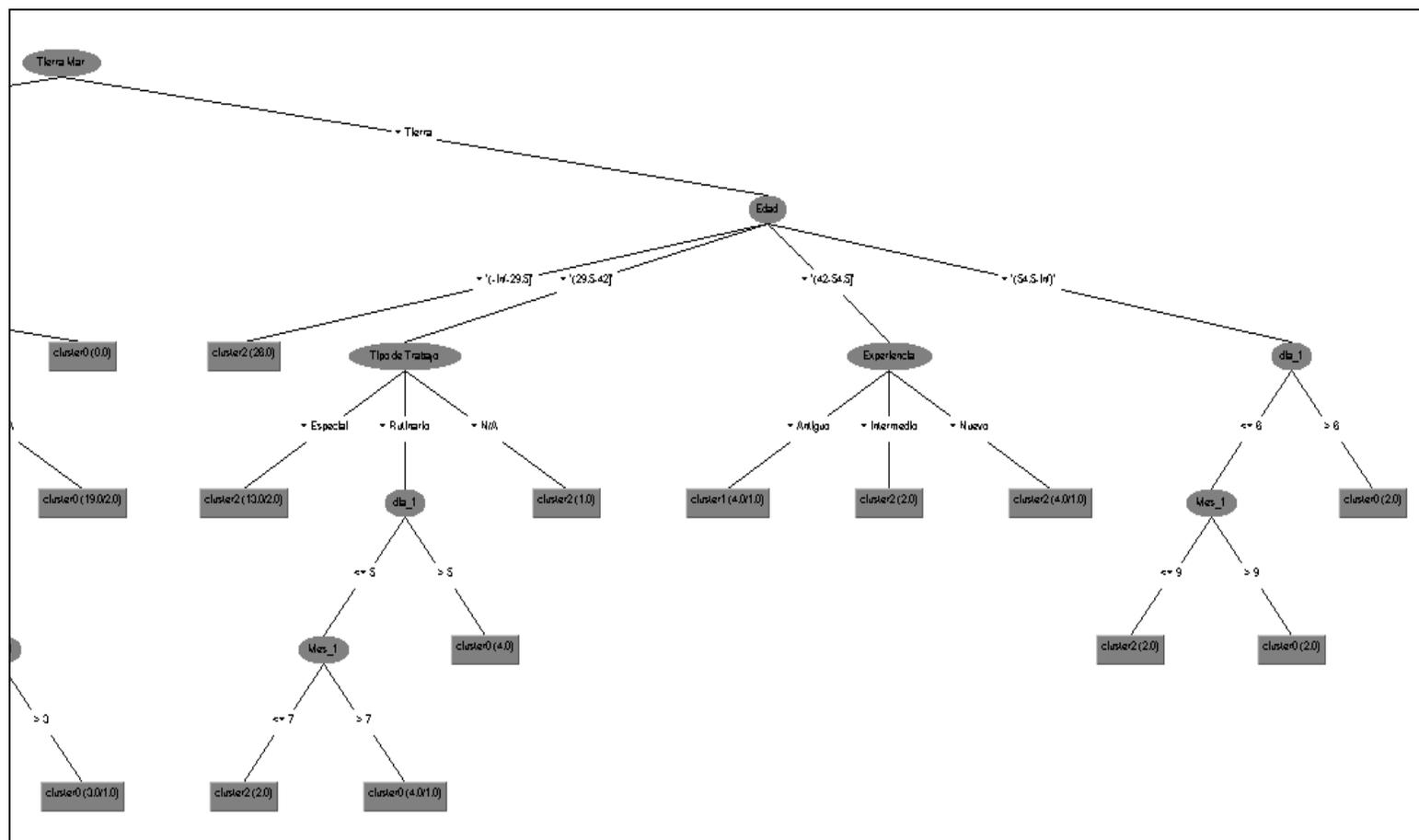
Árbol podado – algoritmo C4.5 (combinación 03)

Figura 3.23. Disgregación del nodo Dot Serv – instancia Servicio



Árbol podado – algoritmo C4.5 (combinación 03)

Figura 3.25. Disgregación del nodo Edad



Árbol podado – algoritmo C4.5 (combinación 03)

Reglas de asociación resultantes en *Weka*Figura 3.26. Reglas de asociación en *Weka*

```

=== Classifier model (full training set) ===

J48 pruned tree
-----

Tierra Mar = Mar
| Dot Serv = Servicio
| | Experiencia = Antiguo
| | | Tipo de Trabajo = Especial: cluster1 (42.0)
| | | Tipo de Trabajo = Rutinario
| | | | Edad = '(-inf-29.5]': cluster0 (3.0)
| | | | Edad = '(29.5-42]':
| | | | | Mes_1 <= 6: cluster1 (5.0)
| | | | | Mes_1 > 6: cluster0 (8.0)
| | | | Edad = '(42-54.5]': cluster1 (14.0/1.0)
| | | | Edad = '(54.5-inf)': cluster0 (4.0/1.0)
| | | Tipo de Trabajo = N/A: cluster1 (24.0)
| | Experiencia = Intermedio
| | | Edad = '(-inf-29.5]': cluster2 (9.0)
| | | Edad = '(29.5-42]': cluster0 (10.0/2.0)
| | | Edad = '(42-54.5]': cluster1 (7.0/1.0)
| | | Edad = '(54.5-inf)': cluster0 (0.0)
| | Experiencia = Nuevo
| | | Tipo de Trabajo = Especial: cluster1 (8.0)
| | | Tipo de Trabajo = Rutinario: cluster0 (6.0/1.0)
| | | Tipo de Trabajo = N/A: cluster1 (1.0)
| Dot Serv = Dotación
| | Tipo de Trabajo = Especial
| | | Edad = '(-inf-29.5]': cluster0 (8.0/2.0)
| | | Edad = '(29.5-42]': cluster0 (18.0/1.0)
| | | Edad = '(42-54.5]': cluster1 (17.0/4.0)
| | | Edad = '(54.5-inf)':
| | | | dia_1 <= 3: cluster1 (4.0)
| | | | dia_1 > 3: cluster0 (3.0/1.0)
| | Tipo de Trabajo = Rutinario: cluster0 (80.0/3.0)
| | Tipo de Trabajo = N/A: cluster0 (19.0/2.0)
| Dot Serv = Especial: cluster0 (0.0)
Tierra Mar = Tierra
| Edad = '(-inf-29.5]': cluster2 (26.0)
| Edad = '(29.5-42]':
| | Tipo de Trabajo = Especial: cluster2 (13.0/2.0)
| | Tipo de Trabajo = Rutinario
| | | dia_1 <= 5
| | | | Mes_1 <= 7: cluster2 (2.0)
| | | | Mes_1 > 7: cluster0 (4.0/1.0)
| | | dia_1 > 5: cluster0 (4.0)
| | Tipo de Trabajo = N/A: cluster2 (1.0)
| Edad = '(42-54.5]':
| | Experiencia = Antiguo: cluster1 (4.0/1.0)
| | Experiencia = Intermedio: cluster2 (2.0)
| | Experiencia = Nuevo: cluster2 (4.0/1.0)
| Edad = '(54.5-inf)':
| | dia_1 <= 6
| | | Mes_1 <= 9: cluster2 (2.0)
| | | Mes_1 > 9: cluster0 (2.0)
| | dia_1 > 6: cluster0 (2.0)

Number of Leaves : 34
Size of the tree: 52

```

Reglas de asociación resultantes del experimento

4. Conclusiones y recomendaciones

Al aplicar tecnología de minería de datos para tratar de obtener patrones de comportamiento de la accidentabilidad laboral de IMI se ha podido obtener interesantes relaciones y características que comparten estos datos de accidentes, con lo que es factible dilucidar nuevas líneas de investigación y estrategias de prevención.

En primer lugar conviene resumir los resultados más concluyentes de los experimentos realizados:

1. La aplicación del algoritmo *K-means* en la tarea de agrupamiento obtiene 03 grupos o patrones de accidentabilidad basados particularmente por trabajos que se realizan en el mar y en cuyo caso ocurren accidentes en trabajos en barcasas y aquellos que se registran en plataformas petroleras; mientras que por otro lado, tenemos un grupo de accidentes que se registran en operaciones en tierra.
2. La aplicación del algoritmo C4.5 en la tarea de clasificación parte de la caracterización de accidentes basados en dos posibles instancias, los que ocurren en el mar y los que se registran en tierra. Se tiene un mayor número de reglas de inducción para el primer caso.
3. Con respecto a los accidentes ocurridos en locaciones marinas se tiene como característica importante los trabajos relacionados con reparación de estructuras en plataformas o actividades relacionadas con plataformas. Así mismo las actividades de maniobra (realizadas por maniobristas) resultan muy significativas en frecuencia de accidentabilidad.
4. Para el caso de accidentes ocurridos en mar y por trabajadores que no pertenecen a la dotación de la barcaza o buque en donde se realiza el trabajo (recuérdese que grupos de trabajadores tales como soldadores, cortadores, entre otros, pueden ser enviados a locaciones marinas a realizar trabajos específicos) se tiene como característica principal que la frecuencia de accidentabilidad para el personal antiguo es muy significativa independientemente del tipo de trabajo a realizar.
5. Los accidentes en trabajos rutinarios caracterizan su frecuencia de ocurrencia en barcasas o buques y del mismo modo sucede en accidentes que no ocurrieron en situación de trabajo. Es particularmente interesante el hecho que la edad no es un factor influyente en este tipo de casos. Como conclusión interesante se tiene la deducción que los trabajos rutinarios en este tipo de locaciones resulta más significativo a nivel de riesgos que aquellos que ocurren en otro tipo, tales como talleres en tierra.
6. El otro grupo muy significativo de ocurrencia de accidentes son aquellos registrados en locaciones tierra; es decir los que ocurren en muelle, talleres u oficinas. En estos casos se caracteriza de forma significativa aquellos que ocurren en trabajos especiales y a personal con edades menores a 30 años. De hecho los trabajos de reparación de motores o acondicionamiento de barcos y barcasas resulta como trabajos de mayor tasa de riesgo.

7. La dimensión temporal de los hechos no parece ser un factor muy influyente en la ocurrencia de accidentes aunque en términos generales se puede decir que en ambos casos se tiene un comportamiento oscilante. Los accidentes parecen incrementarse, luego reducirse y así mantener dicha tendencia. Resulta lógico pensar que pueda ser así en consideración a la respuesta de la empresa ante la ocurrencia de accidentes.

El Departamento de QHSE de IMI validó los resultados de forma positiva confirmando los grupos o categorías obtenidas además de considerar algunos patrones importantes como lo es el hecho de tener una alta incidencia de trabajadores accidentados que no pertenecen a la dotación de las barcasas o buques así como la relevancia de definir estándares en las operaciones de alto riesgo entre barcasas o buques y plataformas petroleras.

El presente trabajo fue entregado y expuesto a las gerencias de la empresa IMI quedando como un importante aporte en el estudio de la accidentabilidad que registra esta empresa.

Anexo C.1 - Listado de Tablas y gráficos

a. Tablas

- ✓ Tabla 2.1. Descripción de atributos seleccionados
- ✓ Tabla 2.2. Nuevo estado de la variable Día
- ✓ Tabla 2.3. Nuevo estado de la variable Mes
- ✓ Tabla 3.1. Iteraciones para encontrar la semilla más adecuada con 3 cluster.
- ✓ Tabla 3.2. Valores más representativos.
- ✓ Tabla 3.3. Selección de atributos.
- ✓ Tabla 3.4. Aplicación del algoritmo C4.5.

b. Figuras

- ✓ Figura 2.1. Distribución de la variable Edad
- ✓ Figura 2.2. Distribución de la variable Cargo
- ✓ Figura 2.3. Distribución de la variable Experiencia
- ✓ Figura 2.4. Distribución de la variable Día
- ✓ Figura 2.5. Distribución de la variable Mes
- ✓ Figura 2.6. Distribución de la variable Turno
- ✓ Figura 2.7. Distribución de la variable Lugar
- ✓ Figura 2.8. Distribución de la variable Dotación-Servicio
- ✓ Figura 2.9. Distribución de la variable Tipo de trabajo
- ✓ Figura 2.10. Discretización de la variable Edad
- ✓ Figura 2.11. *Dataset* cargado en el *Weka*
- ✓ Figura 3.1. Resultado del algoritmo *K-means* en *Weka*
- ✓ Figura 3.2. Distribución de *clusters*
- ✓ Figura 3.3. Interacción de *clusters* con variable edad
- ✓ Figura 3.4. Interacción de *clusters* con variable cargo
- ✓ Figura 3.5. Interacción de *clusters* con variable experiencia
- ✓ Figura 3.6. Interacción de *clusters* con variable día
- ✓ Figura 3.7. Interacción de *clusters* con variable mes
- ✓ Figura 3.8. Interacción de *clusters* con variable turno
- ✓ Figura 3.9. Interacción de *clusters* con variable lugar

- ✓ Figura 3.10. Interacción de *clusters* con variable tierra mar
- ✓ Figura 3.11. Interacción de *clusters* con variable dotación servicio
- ✓ Figura 3.12. Interacción de *clusters* con variable tipo de trabajo
- ✓ Figura 3.13. Interacción de *clusters* con instancias de la variable edad
- ✓ Figura 3.14. Interacción de *clusters* con instancias de la variable cargo
- ✓ Figura 3.15. Interacción de *clusters* con instancias de la variable turno
- ✓ Figura 3.16. Interacción de *clusters* con instancias de la variable lugar
- ✓ Figura 3.17. Interacción de *clusters* con instancias de la variable dotación servicio
- ✓ Figura 3.18. Interacción de *clusters* con instancias de la variable tipo de trabajo
- ✓ Figura 3.19. Interrelación de variables edad y tierra mar
- ✓ Figura 3.20. Interrelación de variables lugar y turno
- ✓ Figura 3.21. Resultados combinación
- ✓ Figura 3.22. Resultados combinación 03
- ✓ Figura 3.23. Disgregación del nodo Dot Serv – instancia Servicio
- ✓ Figura 3.24. Disgregación del nodo Dot Serv – instancia Dotación y Especial
- ✓ Figura 3.25. Disgregación del nodo Edad
- ✓ Figura 3.26. Reglas de asociación en *Weka*

Anexo C.2 - Listado de cargos o puestos de trabajo definidos según organigrama de IMI (instancias posibles de la variable cargo).

1. Arenador
2. Armador
3. Capitán de barcas
4. Capitán de embarcación
5. Carpintero
6. Cortador
7. Despachador
8. Electricista
9. Esmerilador
10. Gasfitero
11. Maniobrista tierra
12. Marinero maniobrista
13. Mecánico
14. Motorista
15. Oficial de cubierta
16. Oficial de puente
17. Oiler
18. Operador de grúa
19. Operario de caldero
20. Pintor
21. Radio operador
22. Soldador
23. Supervisor
24. Tapicero

Anexo C.3 - Listado de locaciones o lugares en donde se realizan las operaciones de IMI (instancias posibles de la variable Lugar).

1. Barcaza tender: puede ser para facilidades de producción o perforación.
2. Barcaza de trabajo: utilizadas para trabajos de mantenimiento, instalación, etc. de plataformas.
3. Baterías: base en tierra para la recepción de los hidrocarburos.
4. Boyas: elemento flotante utilizado en el sistema de anclaje de las barcasas o buques grandes.
5. Buques: categoría que involucra a barcos remolcadores, *suppl* y para maniobra con anclas.
6. Flotante: artefacto naval flotante sin propulsión propia que sirve como soporte logístico de las operaciones marinas ubicadas lejos de puerto o muelle.
7. Lancha de recorrido: barco utilizado para el transporte de pasajeros entre plataformas o puntos específicos en el mar. Estas lanchas navegan en locaciones marinas alejadas de costa o puerto.
8. Muelles: construcción metálica ubicada en la playa y sujeta firmemente al lecho marino que sirve de interface entre tierra y los artefactos navales (propulsados o no) para el soporte de diferente tipo de operaciones (carga y descarga de materiales, embarco de personal, etc.).
9. Oleoducto: conjunto de líneas que sirven para transportar el hidrocarburo de un punto al otro.
10. Plataformas: estructura metálica sujeta firmemente al lecho marino utilizada para la explotación de hidrocarburos en el zócalo continental.
11. Talleres: categoría que involucra a todos los talleres que tiene IMI en tierra ya sea como soporte de su flota de barcos o para realizar servicios específicos en el mar.