# PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## FACULTAD DE CIENCIAS E INGENIERÍA

# Proceso de Descubrimiento de Conocimiento para Predecir el Abandono de Tratamiento en una Entidad de Salud Pública

Tesis para optar por el Título de Ingeniero Informático, que presenta el Bachiller:

**Julio Christians Candela Cáceres**

**ASESOR: Ing. Gissella Bejarano Nicho**

Lima, Abril de 2015

# Resumen

El presente proyecto académico de fin de carrera tiene como objetivo mostrar el proceso automatizado de cada etapa del proceso de descubrimiento con el fin de predecir el abandono en los tratamientos de cáncer de una entidad de salud pública con una precisión eficiente basándose en características o factores determinados en la etapa de análisis junto con los miembros de la institución. La información resultante servirá de apoyo para que los administradores de la entidad de salud puedan plantear las políticas y estrategias personalizadas de retención de pacientes.

Como se mencionó anteriormente, se tomaron en cuenta todas las etapas del proceso de descubrimiento de conocimiento - análisis, extracción, pre-procesamiento, estimación del modelo e interpretación - para que la información resultante pueda ser confiable y oportuna para la toma de decisiones. Adicionalmente, como parte de la etapa de extracción de datos, se encontró la necesidad de diseñar un DataMart que organice y facilite el análisis de información, no solo para el proyecto actual, sino para otras necesidades que puedan surgir en el futuro.

Cada etapa tuvo apoyo de herramientas de software y metodologías que han sido ampliamente usadas con éxito en este tipo de proyectos. Se escogieron herramientas gratuitas que tendrían mayor apoyo a los requerimientos del proyecto como la automatización de los procesos, diseño del DataMart y el proceso general de Minería de Datos.

En conclusión, el proyecto culminó con éxito cumpliendo los estipulado en cada uno de los resultados esperados, por lo cual, se puede determinar que el proceso automatizado podrá ser útil para determinar que pacientes abandonan su tratamiento y brindar la información oportuna a los encargados de tomar las decisiones.

## Tabla de contenido


CAPÍTULO 1 1

**1.1 PROBLEMÁTICA 1**
**1.2 OBJETIVO GENERAL 5**
**1.3 OBJETIVOS ESPECÍFICOS 5**
**1.4 RESULTADOS ESPERADOS 5**
**1.5 HERRAMIENTAS, MÉTODOS, METODOLOGÍAS Y PROCEDIMIENTOS 6**
1.5.1 HERRAMIENTAS 8
1.5.2 MÉTODOS Y PROCEDIMIENTOS 11
1.5.3 METODOLOGÍAS 20
**1.6 ALCANCE 22**
1.6.1 LIMITACIONES 23
1.6.2 RIESGOS 23
**1.7 JUSTIFICATIVA Y VIABILIDAD DEL PROYECTO 26**
1.7.1 JUSTIFICATIVA 26
1.7.2 VIABILIDAD 27

CAPÍTULO 2 29

**2.1 MARCO CONCEPTUAL 29**
2.1.1 OBJETIVO DEL MARCO CONCEPTUAL 29
2.1.2 CONCEPTOS RELACIONADOS CON MINERÍA DE DATOS 29
2.1.3 CONCEPTOS RELACIONADOS CON EL SECTOR SALUD 36
2.1.4 CONCLUSIÓN 38
**2.2 ESTADO DEL ARTE 38**
2.2.1 OBJETIVOS DE LA REVISIÓN DEL ESTADO DEL ARTE 39
2.2.2 MÉTODO USADO EN LA REVISIÓN DEL ESTADO DEL ARTE 39
2.2.3 USO DE TÉCNICAS DE MINERÍA DE DATOS PARA DETERMINAR Y PREDECIR EL PERIODO DE ESTANCIA DE PACIENTES CARDIACOS 39
2.2.4 FACTORES DE DESERCIÓN DE UN PROGRAMA PEDIÁTRICO DE CONTROL DE PESO 41
2.2.5 USO DE MINERÍA DE DATOS PARA DESCRIBIR EL TIEMPO DE ESTANCIA EN UN HOSPITAL 42
2.2.6 APLICACIÓN DE MÁQUINA DE VECTOR DE SOPORTE PARA LA PREDICCIÓN DEL CUMPLIMIENTO DE LA MEDICACIÓN EN PACIENTES CON INSUFICIENCIA CARDIACA 42
2.2.7 CLASIFICACIÓN Y ANÁLISIS SECUENCIAL DE PATRONES PARA MEJORAR LA EFICIENCIA DE GESTIÓN Y PROVEER UN MEJOR SERVICIO MÉDICO EN CENTROS DE SALUD PÚBLICOS 43
2.2.8 FACTORES QUE AFECTAN LA DESERCIÓN TEMPRANA Y EL CURSO DEL TRATAMIENTO TARDÍO DE UN TRATAMIENTO ANTIDEPRESIVO DE DEPRESIÓN EN CIRCUNSTANCIAS NATURALES 44
2.2.9 OTROS TRABAJOS RELACIONADOS 45
2.2.10 CONCLUSIONES SOBRE EL ESTADO DEL ARTE 49

CAPÍTULO 3: CREACIÓN DEL DATAMART DE PACIENTES 50

**3.1 RESULTADO ESPERADO 1: CREACIÓN DEL MODELO DE DATAMART CON LOS DATOS NECESARIOS 50**
3.1.1 MODELO DEL DATAMART 53
3.1.2 DESCRIPCIÓN DE LAS ENTIDADES INFLUYENTES EN EL MODELO 53
3.1.3 TIPO DE DATAMART 54

3.1.4 Estándares del modelo 55
3.1.5 Conclusiones 55
**3.2 Resultado Esperado 2: Proceso ETL que permita la integración automatizada de la información 56**
3.2.1 Carga de tablas maestras 57
3.2.2 Carga de Tablas Transaccionales 58
3.2.3 Conclusiones 58

CAPÍTULO 4: PRE-PROCESAMIENTO DEL MODELO DE MINERÍA DE DATOS 60

**4.1 Resultado Esperado 3: Método de tratamiento de valores nulos o inconsistentes 60**
4.1.1 Eliminación de muestras o variables con exceso de valores nulos 61
4.1.2 Reemplazar los valores nulos con valores probables o esperados 62
4.1.3 Conclusiones 62
**4.2 Resultado Esperado 4: Método de normalización de datos 63**
4.2.1 Transformación de cadena de caracteres 63
4.2.2 Normalización Mínimo-Máximo 64
4.2.3 Suavización 65
4.2.4 Conclusiones 66
**4.3 Resultado Esperado 5: Algoritmo de detección y eliminación de valores anómalos 66**
4.3.1 Algoritmo kNN 67
4.3.2 Pseudocódigo del algoritmo kNN 68
4.3.3 Parámetros 68
4.3.4 Conclusiones 69

CAPÍTULO 5: EVALUACIÓN DE LOS ALGORITMOS 70

**5.1 Resultado Esperado 6: Modelo de datos para cada algoritmo. 70**
5.1.1 Tipo de Archivo 70
5.1.2 Propiedades de cada Conjunto de Datos 72
5.1.3 Análisis de cantidades 72
5.1.4 Conclusiones 73
**5.2 Resultado Esperado 7: Análisis de los algoritmos escogidos de Minería de Datos 73**
5.2.1 Análisis de Algoritmos 74
5.2.2 Herramienta Weka 75
5.2.3 Conclusiones 76
**5.3 Resultado Esperado 8: Evaluación de la precisión de los algoritmos y elección del algoritmo más adecuado 76**
5.3.1 Evaluación de los algoritmos por el método de validación cruzada 77
5.3.2 Elección del algoritmo más eficaz 77
5.3.3 Evaluación de desempeño en la normalización 79
5.3.4 Evaluación del desempeño del algoritmo de análisis de valores anómalos 80
5.3.5 Conclusiones 81

CAPÍTULO 6: AUTOMATIZACIÓN DEL PROCESO DE MINERÍA DE DATOS 82

**5.4 RESULTADO ESPERADO 9: PROGRAMA EJECUTABLE PARA LA ACTUALIZACIÓN DEL PROCESO DE MINERÍA DE DATOS 82**
5.4.1 PROCESO AUTOMATIZADO DE MINERÍA DE DATOS 82
5.4.2 VENTANA DE CONFIGURACIÓN DE PARÁMETROS 84
5.4.3 CONCLUSIONES 85
**5.5 RESULTADO ESPERADO 10: REPORTE CON LOS RESULTADOS PARA CADA PACIENTE 86**
5.5.1 CONCLUSIONES 87

6 DISCUSIÓN DE RESULTADOS 88

**6.1 OBJETIVO ESPECÍFICO 1 88**
**6.2 OBJETIVO ESPECÍFICO 2 89**
**6.3 OBJETIVO ESPECÍFICO 3 89**
**6.4 OBJETIVO ESPECÍFICO 4 90**

7 CONCLUSIONES 92

8 REFERENCIAS BIBLIOGRÁFICAS 94

9 ANEXOS 98

**9.1 PROCESO GENERAL DEL PROYECTO PARA EL INEN Y CATÓLICA 98**
**9.2 TABLA: DICCIONARIO DE DATOS DEL MODELO DEL DATAMART 98**
**9.3 DESCRIPCIÓN DE LOS PRINCIPALES PROCESOS ETL PARA LA CARGA DEL DATAMART 104**
**9.4 MÉTODOS DE NORMALIZACIÓN PARA CADA VARIABLE 106**
**9.5 DESCRIPCIÓN DE LOS PRINCIPALES PROCESOS ETL PARA EL PROCESO AUTOMATIZADO DE MINERÍA DE DATOS 107**

## Índice de figuras

Figura 1.1: Detección de outliers _____ 15
Figura 1.2: Algoritmo SVM-Comparación de espacios entre límites _____ 17
Figura 1.3: Árbol de decisión para los atributos X e Y _____ 18
Figura 1.4: Neurona Artificial simple _____ 19
Figura 1.5: Clasificador kNN _____ 19
Figura 1.6: Proceso de Minería de Datos _____ 20
Figura 2.1: Pasos que constituyen el proceso de descubrimiento de conocimiento _____ 30
Figura 2.2: Máquina de aprendizaje _____ 32
Figura 2.3: Interpretación gráfica de Clasificación _____ 33
Figura 2.4: Interpretación gráfica de Regresión _____ 34
Figura 2.5: Interpretación gráfica de Clustering _____ 34
Figura 2.6: Asociación _____ 35
Figura 2.7: Interpretación gráfica de Summarization _____ 35
Figura 2.8: Interpretación gráfica de Modelo de dependencia _____ 35
Figura 2.9: Interpretación gráfica de Detección de cambio y desviación _____ 36
Figura 3.1: Modelo del DataMart _____ 52
Figura 3.2: Proceso ETL general _____ 57
Figura 3.3: Transformación de la tabla maestra “DMT_Diagnostico_MAE” _ 58
Figura 4.1: Ejemplo de normalización Min-Max _____ 65
Figura 4.2: Pseudocódigo del algoritmo kNN _____ 68
Figura 5.1: Estructura del archivo de tipo Arff para el proyecto _____ 71
Figura 5.2: Matriz de Confusión para cada algoritmo _____ 78
Figura 6.1: Flujo general del proceso automatizado de Minería de Datos __ 83
Figura 6.2: Proceso automatizado para ejecutar el algoritmo J48 _____ 83
Figura 6.3: Ventana de Configuración de parámetros del modelo _____ 84
Figura 6.4: Ventana de Configuración de parámetros del algoritmo de detección de valores anómalos _____ 85
Figura 6.5: Ejemplo de reporte de abandono de tratamientos por departamento de atención _____ 87

## Índice de tablas

Tabla 1.1: Cuadro de resultados esperados con sus respectivas Herramientas, Métodos y Procedimientos _____ 7
Tabla 1.2: Cuadro de Riesgos del proyecto _____ 24
Tabla 2.1: Cuadro de Comparación del Estado del Arte _____ 47
Tabla 4.1: Cuadro de resultados de valores nulos _____ 61
Tabla 4.2: Cuadro de variables de cadena de caracteres _____ 64
Tabla 5.1: Cuadro de variables del modelo _____ 72
Tabla 5.2: Cuadro de cantidades por clase _____ 73
Tabla 5.3: Ventajas y Desventajas de los algoritmos elegidos _____ 75
Tabla 5.4: Resultados del método de validación cruzada _____ 77
Tabla 5.5: Sensibilidad y Especificidad para cada algoritmo _____ 79
Tabla 5.6: Curvas de ROC para cada algoritmo _____ 79
Tabla 5.7: Resultados de desempeño-Normalización _____ 80
Tabla 5.8: Resultados de desempeño-Análisis de Outliers _____ 81

# CAPÍTULO 1

## 1.1 Problemática

Hoy en día, el uso de nuevas herramientas tecnológicas de tratamiento de información se ha convertido en una necesidad para el sector salud debido a que permite generar procesos más eficientes y sirve de apoyo para la toma de decisiones. Los sobrecostos generados en una institución de salud se deben, comúnmente, a procesos ineficaces, errores, corrupción y el poco conocimiento acerca de cómo brindar un buen tratamiento a determinado tipo de paciente [HOR 2007]. Para poder encontrar una solución a estos problemas se requiere de una correcta planificación apoyada por algún proceso de explotación de la información [HOR 2007].

Diariamente, en un instituto de salud público, se generan grandes cantidades de información de acuerdo a los servicios que son brindados. Dicha información es considerada de un nivel complejo alto debido a su gran volumen, sus distintas fuentes de origen, el grado de veracidad de sus fuentes, entre otros. Estos factores pueden originar que los datos sean almacenados de forma dispersa y con mucha información irrelevante, lo cual obstaculiza la labor de los interesados en el análisis e investigación. Por esta razón, un reto en el sector de salud pública es enfrentarse al hecho de cómo aprovechar mejor la información que se tiene disponible [OLA 2014].

En los últimos años, el Ministerio de Salud del Perú (MINSA) ha realizado políticas para mejorar y fortalecer los sistemas de información de salud en el Perú. Muchos hospitales han digitalizado los datos de los pacientes y utilizan bases de datos para almacenarlos, lo cual, por ejemplo, hace la labor de recolección de historias clínicas más rápida; sin embargo, aún no es capaz de brindar apoyo para un análisis profundo que permita tomar decisiones certeras. Según la conferencia realizada por la Oficina General de Estadística e Informática (OGEI) del MINSA y la Organización Panamericana de la Salud [MIN1 2011], “Una información objetiva, oportuna y confiable es vital para las decisiones en salud pública y es la base para las políticas, programas, presupuestos y evaluaciones a nivel de gobierno”. Con respecto a este último punto, se puede añadir la importancia de que dicha información sirva para generar conocimiento, no solo para resolver problemas actuales, sino también, que permita tener un grado de flexibilidad capaz de adaptarse por sí mismo a necesidades

futuras de información y análisis. Esta característica hace valiosa a la información y permite la optimización de recursos en un sector crítico como es el de la salud pública en el Perú.

De acuerdo al plan estratégico del instituto nacional de salud [INS 2012], se ha fijado como objetivo estratégico el monitorio, supervisión, evaluación y control de programas para varias enfermedades, dentro de los cuales se contempla el interés por reducir problemas que puedan surgir durante el tratamiento, como es el abandono. Por lo tanto, una preocupación latente para la mayoría de institutos de salud es poder predecir el comportamiento de cada paciente con respecto al tratamiento que debe seguir. Cabe mencionar que se considera tratamiento al proceso que sigue un paciente desde que se le detecta la enfermedad hasta que deja de asistir a las consultas por dicha enfermedad, ya sea por cura de la enfermedad, voluntad propia o fallecimiento. La falta de este conocimiento dificulta que se pueda realizar una mejor planificación en lo que se refiere a destinar recursos dependiendo del tipo de tratamiento, lo cual, finalmente, produce gastos innecesarios y procesos inapropiados que no colaboran con una buena gestión de salud. Además, poder obtener este conocimiento daría la posibilidad de brindar un servicio personalizado al paciente, y así, aumentar la calidad de la salud pública.

Dentro de este marco, se puede identificar dos problemas o necesidades que se deben cubrir en el sector salud: el abandono del tratamiento y la incertidumbre acerca de la duración del tratamiento.

Según el estudio de cohorte para la tuberculosis realizado por el MINSA [MIN2 2011], se identificó un aumento considerable en la tasa de abandono del tratamiento desde el 2006 hasta el 2010, lo cual representa un desafío para combatir esta enfermedad. La deserción en los tratamientos de enfermedades críticas, como la tuberculosis, tiene graves consecuencias para los pacientes, como es el desarrollo de la enfermedad el cual puede generar un peligro de muerte, y para el entorno al cual se encuentra expuesto, como es la proliferación y fortalecimiento de la enfermedad. Además, se genera una gran pérdida de dinero proveniente de los fondos del Estado. Así como para la tuberculosis, el Estado también destina una fuerte inversión para tratamientos de cáncer debido al grado crítico de esta enfermedad. De acuerdo con el plan nacional para el fortalecimiento de la prevención y control del cáncer en el Perú [INE 2006], cada tratamiento de cáncer representa un promedio de $20,000 dólares, lo cual, en caso se produjera un abandono del tratamiento, sería una gran cantidad de dinero

perdido difícil de reinvertir en otros pacientes. Por esa razón, se han realizado algunas políticas con el fin de atacar esta situación de forma prioritaria dependiendo de la enfermedad a tratar; sin embargo, el problema no puede ser resuelto fácilmente ya que se deben que manejar grandes cantidades de información. Poseer el conocimiento acerca de quienes abandonan su tratamiento daría las pautas necesarias a seguir para tomar acciones personalizadas basándose en evidencia científica, y así, aumentar la probabilidad de éxito que pueda tener la solución que sea implementada [INS 2011].

En la mayor parte de los institutos de salud pública, los servicios de permanencia y la duración de tratamientos de los pacientes son altos generadores de costos. Incluso, debido a la gran cantidad de pacientes que deben atenderse diariamente, no se tiene una buena distribución del uso de los recursos disponibles para cada tipo de tratamiento. Esto afecta, generalmente, a las personas de bajos recursos, las cuales representan la mayor parte de los pacientes, porque genera insatisfacción y servicios que no tienen la calidad adecuada. Los tratamientos para enfermedades como el cáncer, por ejemplo, requieren de un conjunto de etapas complejas dependiendo de la fase y el tipo de cáncer que se le diagnostique al paciente. La dificultad para poder identificar estas características o patrones en el tiempo del tratamiento es una barrera para conseguir el conocimiento que nos permita realizar una buena planificación de los recursos [HAC 2013].

Una solución para manejar grandes volúmenes de datos y obtener un valor agregado, al cual llamamos conocimiento, es la Minería de Datos. La Minería de Datos, la cual forma parte del proceso de descubrimiento de conocimiento, es un proceso iterativo que permite el descubrimiento de patrones, conclusiones e información derivada de un conjunto de datos basándose en el aprendizaje para poder generar nuevo conocimiento. Dentro de este proceso se puede destacar una de sus tareas: creación de modelos predictivos, los cuales servirán de apoyo para la gestión de la salud pública [MEH 2011]. Muchos centros de salud internacionales se han percatado de su utilidad como una herramienta para la resolución de problemas y evitar el exceso de costos. La parte fundamental de un proceso de Minería de Datos es la recolección de datos; por esta razón, los datos acerca de los pacientes, tratamientos, doctores, entre otros, deben estar correctamente almacenados en una base de datos confiable para que no degraden los resultados del proceso [HOR 2007]. Como se mencionó anteriormente, muchos hospitales públicos ya tienen los datos de los pacientes digitalizados, lo cual es de gran ayuda para tener éxito en el desarrollo de esta

herramienta; sin embargo, aún se debe realizar una depuración de estos para obtener una mayor precisión. El resultado final de este proceso servirá de guía para tomar decisiones en base a lo que pueda ocurrir en el futuro.

Por todo lo ya explicado, el proyecto de fin de carrera tiene como objetivo la automatización de un proceso de descubrimiento de conocimiento que permita predecir el comportamiento de un paciente con respecto a la continuidad en su tratamiento para que se puedan emplear mecanismos de incentivo con tal que dichos pacientes puedan retomar sus respectivos tratamientos y sus enfermedades no se desarrollen provocando graves consecuencias. Esto se logrará a través de la revisión de casos pasados para determinar las causas principales, y así predecir el comportamiento de los nuevos pacientes basándonos en ese conocimiento adquirido. Los institutos de salud tendrán la posibilidad de brindar un servicio orientado al paciente, demostrando su preocupación hacia ellos, y realizar las acciones necesarias para reducir la frecuencia de abandono del tratamiento. Los administradores de la entidad de salud tendrán la posibilidad de contar con este conocimiento periódicamente. Finalmente, los costos generados por la implementación del modelo predictivo son mínimos, lo cual la convierte en una solución viable y atractiva para atacar el problema del abandono del tratamiento.

## 1.2 Objetivo general

Automatizar un proceso de descubrimiento de conocimiento para una institución de salud pública que permita determinar el comportamiento de los pacientes con respecto a la continuidad en sus tratamientos.

## 1.3 Objetivos específicos

1) Diseñar un modelo de base de datos que integre las diferentes fuentes de información de la entidad de salud, con el fin de tener un conjunto de datos de entrada organizado y fácil de utilizar.

2) Identificar los escenarios donde la variación y anomalías de las características en los registros de pacientes pueda influir de manera negativa en la precisión del modelo, para luego realizar las acciones correspondientes para evitar que afecten los resultados.

3) Evaluar los algoritmos del proceso de Minería de Datos a través de un número de iteraciones del conjunto de datos, con el fin de obtener la mayor precisión para nuestro escenario.

4) Automatizar el proceso de Minería de Datos con la carga de nuevos datos, periódicamente, para que se pueda adaptar a condiciones futuras.

## 1.4 Resultados esperados

1) Diseñar un modelo de base de datos que integre las diferentes fuentes de información de la entidad de salud.

    - Modelo de DataMart creado con los datos requeridos.
    - Proceso ETL que permita la integración automatizada de la información.

2) Identificar los escenarios donde la variación y anomalías de las características en los registros de pacientes pueda influir de manera negativa en la precisión del modelo.

   - Método de tratamiento de valores nulos implementado.
   - Método de transformación de datos crudos implementado de acuerdo a los requerimientos de cada modelo de clasificación.
   - Algoritmo de detección de valores anómalos y su posterior eliminación del modelo en caso sea necesario.

3) Evaluar los algoritmos del proceso de Minería de Datos a través de un número de iteraciones del conjunto de datos.

   - Modelo de datos (conjunto de entrenamiento y conjunto de prueba) para cada iteración del algoritmo.
   - Análisis de los algoritmos, utilizando las bibliotecas existentes.
   - Evaluación de la precisión de los algoritmos y elección del algoritmo más adecuado (aproximadamente, la precisión aceptable debe ser mayor a 60%, de acuerdo con las investigaciones del Estado del Arte).

4) Automatizar el proceso de Minería de Datos con la carga de nuevos datos, periódicamente, para que se pueda adaptar a condiciones futuras.

   - Programa ejecutable para la actualización del proceso de Minería de Datos
   - Informe de reportes periódicos con los resultados para cada paciente.

## 1.5 Herramientas, métodos, metodologías y procedimientos

En la siguiente tabla se muestran las herramientas, métodos y procedimientos que permitirán lograr los resultados esperados, los cuales fueron mencionados anteriormente.

**Tabla 1.1**. Cuadro de resultados esperados con sus respectivas Herramientas, Métodos y Procedimientos. Elaboración propia.

| Resultado Esperado | Herramientas, Métodos y Procedimientos |
|---|---|
| R1) DataMart creado con los datos requeridos. | • DataWarehouse y DataMart: Se aplicarán las técnicas para el diseño del DataMart utilizando como fuente de datos un DataWarehouse.<br>• CA Erwin DataModeler. |
| R2) Proceso ETL que permita la integración automatizada de la información. | • Herramientas de Software: RapidMiner.<br>• Pentaho |
| R3) Método de tratamiento de valores nulos implementado. | • Missing Values: Manejo de valores nulos o eliminación |
| R4) Método de transformación de datos crudos implementado de acuerdo a los requerimientos de cada modelo de clasificación. | • Transformación de datos crudos |
| R5) Algoritmo de detección de outliers y su posterior eliminación del modelo en caso sea necesario. | • Outliers: Detección y eliminación |
| R6) Modelo de datos para cada entrada del algoritmo. | • Conjunto de Entrenamiento y Conjunto de Prueba: Se aplicarán las técnicas para la entrada de cada algoritmo. |
| R7) Análisis e implementación de los algoritmos, utilizando las bibliotecas existentes. | • Herramientas de Software: Weka, JDM API, Rapid Miner, Rhadoop, Orange y SAS. Se escogerá la que mejor se adecue a las necesidades del proyecto.<br>• Algoritmos de clasificación (SVM, Árboles de Decisión, Redes Neuronales y kNN) |
| R8) Evaluación de los algoritmos utilizando el método de validación cruzada (aproximadamente, la precisión aceptable debe ser mayor a 60%, de acuerdo con las investigaciones del Estado del Arte). | • Algoritmos de clasificación (SVM, Árboles de Decisión, Redes Neuronales y kNN): Comparación de medias para determinar los mejores y se utilizará el método de validación cruzada para los conjuntos de datos. |
| R9) Informe de los resultados de precisión de los algoritmos, así como la respectiva interpretación | • Herramientas de Software: Weka. |
| R10) Programa ejecutable para la actualización del proceso de Minería de Datos | • Herramientas de Software: RapidMiner, Pentaho. |

| | |
|---|---|
| R11) Informe de reportes periódicos con los resultados para cada paciente. | • Herramientas de Software: Pentaho, RapidMiner, Orange y R Hadoop. Se escogerá la que mejor se adecue a las necesidades del proyecto. |

### 1.5.1 Herramientas

En la actualidad, la Minería de Datos se ha convertido en una de las técnicas con más aplicaciones en los diferentes sectores. Esto ha dado origen a una gran cantidad de herramientas de software que permiten facilitar la complejidad del proceso en general. Algunas de estas son de código abierto y otras bajo licencia:

**JDM API**

JDM API es una herramienta basada en JDM, el estándar de Java para la Minería de Datos. Permite utilizar los lenguajes PL/SQL y SQL para la Minería de Datos en Oracle [JDM 2014]. JDM es un estándar, en java puro, que permite brindar arquitecturas de aplicación flexibles que se puedan adaptar a las soluciones que plantean los usuarios del negocio. Las principales características de JDM consisten en: [HOR 2007]

- Independencia del proveedor
- Soluciones de múltiples proveedores
- Mayor interoperabilidad

La base de datos de Oracle provee una gran cantidad de funciones y algoritmos de Minería de Datos. Actualmente, se encuentra en la versión JDM 1.1, la cual incluye nuevas funcionalidades como integración y transformación de los datos, modelos lineales generalizados, entre otros [JDM 2014].

**RapidMiner**

RapidMiner es una de las herramientas líderes en plataformas de análisis avanzado. Su diseño visual intuitivo y la poca necesidad de conocimiento de programación convierten a RapidMiner en una de las herramientas preferidas por los analistas del negocio. Cuenta con una versión gratuita y otras cuatro bajo licencia cuyos precios varían de acuerdo a las propiedades que brinda, como son procesamiento, soporte

técnico, sistemas de base de datos, etc. Ente las principales ventajas que tiene esta herramienta, alineadas con el proyecto, están [RAP 2014]:

- Tiene integrado diferentes métodos de integración, transformación y visualización de datos.
- Permite integrar algoritmos propios dentro de RapidMiner ya que posee API's de código abierto.
- Permite obtener visualizaciones potentes de los resultados obtenidos.
- Permite trabajar con grandes conjuntos de datos, a diferencia de muchas otras herramientas.
- Puede ser ejecutado en todos los sistemas operativos.

**SAS**

SAS es una herramienta orientada a las soluciones de Minería de Datos para grandes empresas, por lo cual su uso es bajo licencia. SAS provee al analista modelos predictivos y descriptivos de alta precisión que puedan soportar la cantidad de datos de las grandes empresas. Entre sus beneficios, alineados con el proyecto, se encuentran [SAS 2014]:

- Apoyo completo en el proceso de Minería de Datos (integración, pre procesamiento, algoritmo, resultados).
- Permite trabajar con grandes volúmenes de datos.
- Permite generar modelos de Minería de Datos de manera sencilla.
- Provee una alta precisión, lo cual aumenta la toma de decisiones acertada.
- Posee una plantilla de modelos para los principales problemas de negocio.

**Weka**

Weka es una herramienta de Minería de Datos de código abierto hecha en Java por miembros de la Universidad Waikato. Es una colección de algoritmos que apoyan las tareas de Minería de Datos, especialmente, máquinas de aprendizaje. Entre sus principales características, alineadas con el proyecto, destacan [WEK 2014]:

- Permite trabajar directamente con el conjunto de datos o utilizar la librería de Weka a través de un proyecto en Java.

- Provee herramientas para cada etapa del proceso de Minería de Datos.
- Permite el desarrollo e integración de nuevas máquinas de aprendizaje.

**R Hadoop**

R Hadoop es una herramienta de código abierto, disponible para Windows, Linux y Mac, la cual provee una gran variedad de tareas de Minería de Datos como clasificación, clusterización, modelado lineal y no lineal, entre otros, y técnicas gráficas para desplegar los resultados. Entre sus principales características, alineadas con el proyecto, destacan [RHA 2014]:

- Otorga facilidad para manejar y almacenar los datos.
- Provee una gran colección integrada de tareas para el análisis de datos.
- Provee herramientas gráficas fáciles de implementar para su posterior análisis.
- Posee un lenguaje de programación, S, simple y bien desarrollado que soporta las tareas de Minería de Datos.

**CA Erwin DataModeler**

Erwin es una de las herramientas de modelado de datos más utilizadas a nivel mundial gracias a su interfaz intuitiva y gráfica. Entre las principales características que provee esta herramienta se encuentran [ERW 2014]:

- Apoya la administración de complejas estructuras de datos.
- Facilita la creación de la base de datos directamente desde el modelo.
- Provee herramientas para la integración e intercambio de metadata con otras herramientas.
- Soporta varios servidores de base de datos.

**Pentaho**

Pentaho es un conjunto de software open source que proveen herramientas para facilitar la integración de datos y la inteligencia de negocios. Su principal característica es que permita desarrollar soluciones basadas en procesos y monitorear el rendimiento de dichos procesos de una forma eficiente. Entre sus principales

características, alineadas con los resultados esperados del proyecto, destacan [PEN 2014]:

- Posee una interfaz gráfica intuitiva y herramientas fáciles de aprender.
- Permite realizar el proceso de ETL como una secuencia de procesos.
- Facilita la integración con diferentes fuentes de datos como hojas de cálculo, base de datos, archivos arff, etc.
- Permite utilizar la herramienta de Minería de Datos de Weka.

### 1.5.2 Métodos y Procedimientos

**Data Warehouse**

Un Data Warehouse es un repositorio con información integrada y relevante que permite facilitar la tarea de apoyo a la toma de decisiones estratégicas. La función principal de un Data Warehouse es almacenar la información histórica de una organización de una forma integrada, de tal manera que pueda representar las diferentes estructuras que conforman dicha organización. El resultado final de todo lo almacenado tiene el objetivo de proveer información útil a los usuarios encargados de tomar las decisiones de la organización en general. La ventaja adicional de su implementación es que permitirá descubrir información oculta o difícil de extraer que será muy importante para generar un valor agregado al negocio [MEH 2011].

Según [MEH 2011], el diseño de un Data Warehouse está basado, principalmente en las siguientes características: la clasificación de los datos que serán almacenados en el Data Warehouse y el conjunto de transformaciones que se realizarán para alimentarlo de información. Estas características deben ser planeadas correctamente para que el Data Warehouse sea verdaderamente útil al momento de ser consultado. El proceso de implementación de un Data Warehouse se puede resumir en los siguientes pasos:

a) Identificar los procesos del negocio, sus requerimientos y las decisiones que se toman con respecto a ellos
b) Establecer los requerimientos y crear el modelo de datos. Esto debe ser pensado con tal que puedan adaptarse a las decisiones de cada proceso.

c) Implementar el Data Warehouse en todos los procesos involucrados. Se requiere que los usuarios puedan obtener el mejor provecho de él.

d) Supervisar el desarrollo del Data Warehouse y realizar modificaciones en caso los procesos lo requieran.

En caso la entidad de salud pública, con la cual se desarrollará el proyecto, cuente con un Data Warehouse, este será utilizado como fuente de datos para la creación de un DataMart, el cual sigue, aproximadamente, el mismo proceso de implementación.

**DataMart**

Un DataMart es un conjunto de datos que ha sido diseñado para cubrir necesidades de un área en específico dentro de una organización. En muchas ocasiones, los datos de un DataMart provienen de un Data Warehouse, el cual, al estar formado por una gran cantidad de registros, no es el adecuado para un análisis con fines particulares. El tamaño de datos de un DataMart dependerá del área y el alcance del problema que se plantea resolver [MEH 2011].

Según [MEH 2011], con respecto a la Minería de Datos, un DataMart es de gran utilidad debido a que reduce, o incluso elimina, las tareas que se desarrollan en la etapa de pre-procesamiento (una de las etapas que demandan más tiempo en el proceso de Minería de Datos). Además, sirve de apoyo para poder identificar las características o variables que serán tomadas en cuenta para la implementación y ejecución del modelo predictivo [HOR 2007]. Por esta razón, se deben planear correctamente las tareas de limpieza y transformación de datos antes de la creación del DataMart con el fin de hacer simple el proceso en general. Estas tareas pueden incluir: Transformación de datos crudos, nulos y dependientes del tiempo, así como, la eliminación de “outliers”, los cuales están relacionados directamente con la Minería de Datos.

**Tipos de modelos de DataMart**

- **Esquema Estrella:** Su principal ventaja es que permite un alto rendimiento al momento de realizar consultas. Además posee una alta granularidad debido a su estructura de estrella donde se puede encontrar la tabla principal de hechos y cada punta que representa una dimensión [IBM 2005].

- **Esquema Copo de Nieve:** Su principal ventaja radica en que provee una mayor normalización de las tablas del modelo de datos. Generalmente, la expansión de las dimensiones en un esquema de estrella termina en la implementación de un esquema de copo de nieve. Es utilizado para modelos de base de datos muy complejos, por lo cual su tiempo de consulta es mayor que un esquema de estrella [IBM 2005].

- **Esquema Híbrido:** Consiste en combinar las ventajas de rapidez del esquema de estrella con la alta normalización que posee el esquema de copo de nieve [IBM 2005].

**ETL (Extract-Transform-Load)**

Un proceso ETL es una colección de proceso que permitirá la carga de datos de diferentes fuentes de datos operacionales a un DataWarehouse o DataMart. El proceso general consiste en la extracción de información de dichas fuentes, seguido de un proceso de transformación para obtener nuevas variables o y asegurar la calidad de la información y, finalmente, un proceso de carga al DataMart o DataWarehouse [KIM 2002].

**Missing Values (Valores Perdidos o Nulos)**

Cuando se realizan procesos con grandes cantidades de datos, la probabilidad de tener todos los registros con sus campos completos es relativamente baja. Los valores perdidos son aquellos casos donde el dato no ha sido provisto por el usuario. Esto puede ser generado por omisión del propio sistema, pérdida de datos, entre otros. En caso, un registro o campo presente muchos valores perdidos, se necesita realizar alguna acción para que no afecten a los demás datos [HOR 2007].

Según [MEH 2011], algunos modelos de Minería de Datos, a pesar de tener valores perdidos, pueden procesar correctamente los datos y llegar a un resultado acertado. Sin embargo, el resultado de otros puede ser altamente afectado ya que requieren que todos los valores sean coherentes entre sí. Por esta razón, los valores perdidos deben ser reemplazados durante la etapa de pre procesamiento. Las posibles soluciones se presentan a continuación:

a) Eliminar todas las muestras con valores perdidos: Esta solución es posible cuando estos casos ocurren con poca frecuencia. Por el contrario, si es que son muchos los casos, el conjunto de datos se vería reducido enormemente convirtiéndose en un obstáculo para el aprendizaje del modelo [MEH 2011].
b) Reemplazar el campo vacío con un valor razonable, probable o esperado: Esta solución requiere ser analizada junto con un experto en el área de la cual provienen los datos para llegar a la conclusión de cuál es el valor apropiado. De acuerdo con este enfoque, los valores perdidos podrían ser reemplazados por una constante global, por la media del atributo o por la media del atributo dependiendo de la clase. Reemplazar los valores perdidos con una simple constante obliga a que los casos de distintas clases se homogenicen disminuyendo la precisión del modelo [MEH 2011].
c) Generar un modelo predictivo que sea capaz de encontrar valores probables para los valores perdidos. Para lograr esto, el modelo se basa en las otras características que si están completas y genera el mejor valor para la faltante. En caso el registro presente la mayoría de sus campos con valores perdidos, el modelo no serviría [MEH 2011].

**Transformación de datos crudos**

La transformación de datos crudos es un proceso que permite realizar una posible mejora en los resultados de la Minería de Datos. A continuación se describen algunas técnicas empleadas en esta tarea:

a) Normalización: Consiste en que los valores numéricos o que pueden ser medidos sean escalados a un rango específico para todos. Cuando los valores de los datos crudos tienen mucha diferencia, los que tienen un mayor valor ejercerán más influencia sobre los otros, lo cual, finalmente, distorsionará los resultados del modelo. Entre las principales técnicas de normalización destacan la escala decimal, normalización de mínimos y máximos y la normalización por desviación estándar [MEH 2011].
b) Suavización de datos: Consiste en igualar los valores que tienen mínimas diferencias entre sí sin reducir la calidad de los datos. Por lo tanto, se reduce el número de valores diferentes lo cual es bastante útil para el modelo, especialmente cuando se emplean métodos basados en lógica como los árboles de decisión. Un ejemplo de esta técnica es el redondeo de valores reales [MEH 2011].

c) Diferencias y Relaciones: Consiste en realizar transformaciones menores en las características de entrada o salida que puedan llevar a mejores resultados. Reducir el número de posibles valores para una característica de salida, por ejemplo, mejora le eficiencia del algoritmo que se utilice. Con respecto a las características de entrada, se pueden mezclar dos que formen una nueva que pueda describir mejor el modelo [MEH 2011].

**Outliers (Valores atípicos)**

Según [MEH 2011], los valores atípicos son aquellos que no guardan coherencia con el comportamiento general del modelo. Estos pueden ser causados, principalmente, por errores de medición o por un comportamiento anómalo de la variable. De acuerdo con [HOR 2007], los conceptos de error y outlier están separados debido a que se puede actuar de manera diferente dependiendo del problema. Por ejemplo, un error puede ser reemplazado por un valor correcto; por el otro lado, un outlier debe permanecer con su propio valor y se decidirá a través de un análisis si servirá o no para el modelo. Por esta razón, se debe identificar qué valores pertenecen al grupo de errores y al de los outliers para definir parámetros que nos permitan trabajar con estos valores.

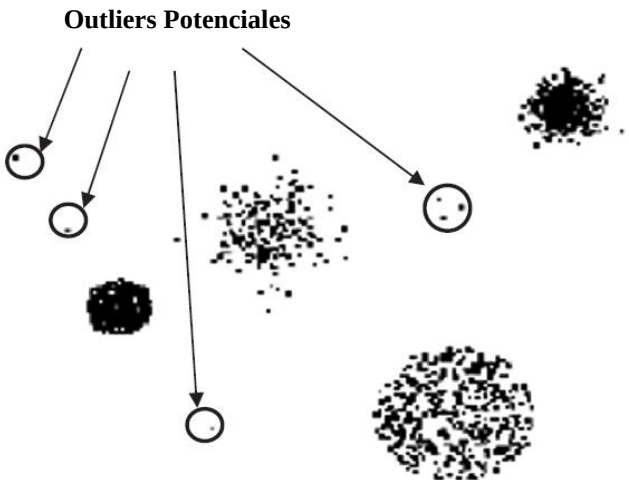


**Figura 1.1.** Detección de outliers. Imagen adaptada de Mehmed Kantardzic

El análisis de outliers debe ser muy minucioso al momento de decidir si una variable es eliminada ya que se podría generar la pérdida de información importante. El objetivo de la detección de outliers es encontrar un número de muestras que muestran

diferencias con el modelo general y determinar si estas deben ser removidas del conjunto de datos [MEH 2011]. Los principales enfoques son los siguientes:

a) Técnicas gráficas o de visualización: Este método es útil cuando el número de dimensiones del modelo es relativamente menor, aproximadamente, tres dimensiones [MEH 2011].
b) Técnicas basadas en estadística: Este método identifica las muestras que se desvían de los supuestos grupos normales a través de métodos estadísticos, utilizando la media y la desviación estándar. Cuando el conjunto de datos tiene una alta cantidad de dimensiones, este método se vuelve inapropiado si es que no se tiene un conocimiento previo de la distribución de los datos [MEH 2011].
c) Técnicas basadas en la distancia: Este método consiste en identificar las muestras que no tenga vecinos cercanos, las cuales serán marcadas como outliers. Para esto, se establece un límite para considerarlo cercano y se calculan las distancias para todas las otras muestras. Es un método simple de implementar y es aplicable a conjunto de datos de varias dimensiones; sin embargo, su complejidad computacional es muy alta y depende de la cantidad de muestras que se pretende analizar [MEH 2011].
d) Técnicas basadas en el modelo: Este método consiste en definir un patrón en las características de un conjunto de datos o clase y marcar como outliers todas aquellas muestras que se desvían. Una opción es definir grupos de muestras cercanas de gran densidad y los grupos que tengan una o pocas muestras serán considerados outliers [MEH 2011].

**Conjunto de Entrenamiento y Prueba**

El conjunto de entrenamiento es un conjunto de datos cuyas clases son conocidas y servirán para alimentar el modelo en la etapa de construcción; por lo tanto, son los encargados de proveer las reglas de clasificación [HAN 2006].

El conjunto de prueba es usado para calcular la precisión del modelo; es decir, comprobar que las reglas de clasificación se cumplan [HAN 2006].

**Validación Cruzada**

En el proyecto, se pretende utilizar el método de validación cruzada para dividir el conjunto de datos inicial en entrenamiento y prueba. Este método consiste en dividir

aleatoriamente el conjunto de datos en “k” subconjuntos de igual tamaño [HAN 2006]. El valor de “k” puede ser de 5 a 10, de acuerdo a los estudios observados en el estado del arte. Para cada iteración del algoritmo, se escogerá un subconjunto diferente a los anteriores que corresponderá al conjunto de prueba y el resto (“k-1” subconjuntos) serán de entrenamiento. Por lo tanto, cada algoritmo escogido para el proceso se ejecutará “k” veces y se conseguirá la precisión promedio, la cual será comparada con la precisión de los otros algoritmos a través de comparación de medias.

**Algoritmos de Clasificación**

Los siguientes algoritmos, cuya elección ha sido basada en estudios pasados vistos en el estado del arte, serán utilizados durante el proceso de Minería de Datos del proyecto académico:

a) Support Vector Machine (SVM): Los principios del algoritmo SVM fueron desarrollados por Vladimir Vapnik. Su diseño permite ser utilizado para problemas de clasificación y regresión, obteniendo buenos resultados incluso cuando el conjunto de entrenamiento es relativamente pequeño. El SVM consiste en determinar el espacio máximo utilizando vectores que puedan forman los límites entre las distintas clases, con el fin de diferenciar las clases para la clasificación de muestras nuevas. Aplicado a muestras de varias dimensiones, el algoritmo permitirá encontrar el hiperplano con el espacio máximo, que pueda separar las distintas clases. El modelo requiere que los atributos sean representados como datos numéricos. La ventaja del algoritmo, con respecto al proyecto, es que permite obtener los factores más influyentes de la decisión de predicción y ha mostrado buenos resultados en las investigaciones pasadas [MEH 2011].

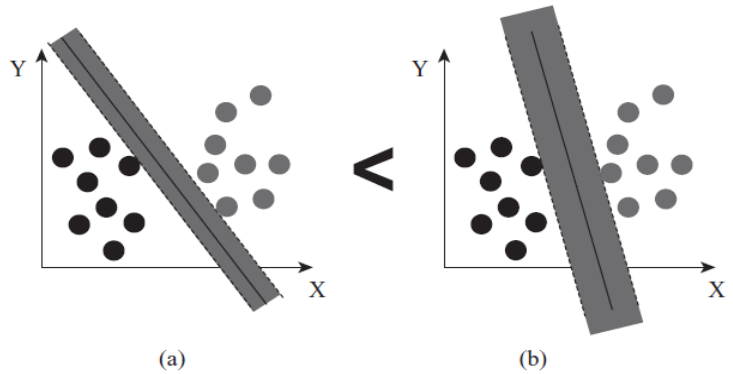


**Figura 1.2.** Algoritmo SVM-Comparación de espacios entre límites. Imagen recuperada de Mehmed Kantardzic.

b) Árboles de Decisión: Es uno de los métodos de clasificación y regresión más utilizados, el cual está basado en un modelo jerárquico donde se realiza una búsqueda hacia abajo para encontrar una solución. Para el proceso de Minería de Datos, un árbol consta de nodos, donde cada atributo es probado, y ramas, las cuales representan el resultado de la prueba. El algoritmo más conocido de árboles de decisión es C4.5. Este usa métodos greedy o golosos para buscar modelos dentro del conjunto de prueba que puedan formar la estructura del árbol. Su diseño permitirá encontrar factores influyentes en la decisión de clasificación [MEH 2011].

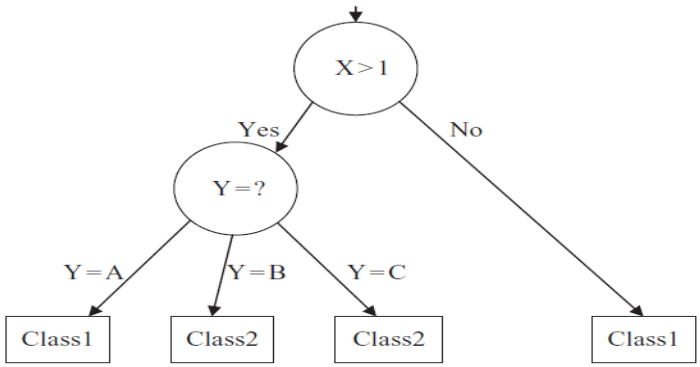


**Figura 1.3.** Árbol de decisión para los atributos X e Y. Imagen recuperada de Mehmed Kantardzic.

c) Redes Artificiales Neuronales (ANT): Este método está basado en la forma como el cerebro humano realiza cálculos a diferencia de la computadora. Las redes neuronales están compuestas por neuronas, las cuales representan unidades de procesamiento, e interconexiones, las cuales representan las relaciones que existen entre neuronas o nodos. Las conexiones entre neuronas permiten que se pueda obtener conocimiento de ellas y este pueda ser generalizado para formar el modelo. La ventaja de las redes neuronales es que puede ser adaptable a las condiciones del conjunto de datos, lo cual otorga una mayor confianza al modelo cuando los datos no han sido completamente procesados [MEH 2011].

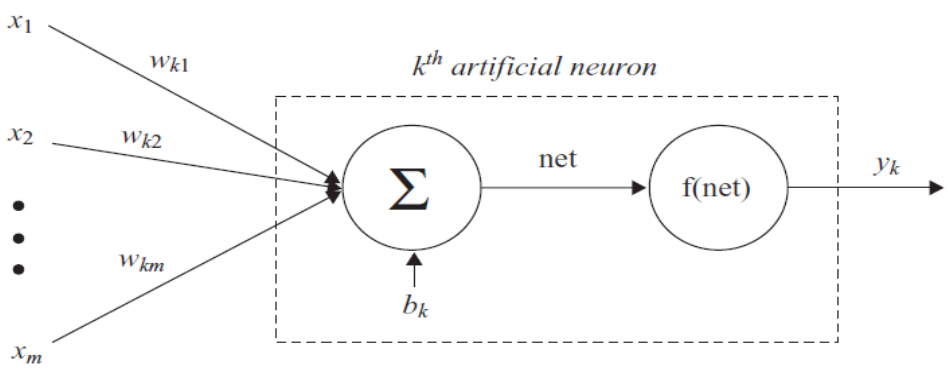


**Figura 1.4.** Neurona Artificial simple. Imagen recuperada de Mehmed Kantardzic.

d) Clasificación por Vecinos más Cercanos (kNN): El clasificador kNN permite determinar grupos de muestras (las cuales forman un límite de decisión) que se encuentren más cerca de una nueva muestra, y a partir de ello clasificarla. El valor "k" representa el número de vecinos más cercanos que el algoritmo utilizará. El entrenamiento de este tipo de algoritmo consiste en determinar el valor óptimo para "k", tal que la precisión del modelo sea lo más óptima posible para un conjunto de prueba. La ventaja de este modelo es que su construcción es relativamente sencilla, ya que solo requiere el valor "k", el conjunto de entrenamiento y un sistema de medición para el cálculo de las distancias entre muestras. Sin embargo, el modelo requiere que los datos puedan representarse numéricamente, lo cual puede complicar un poco el modelo [MEH 2011].

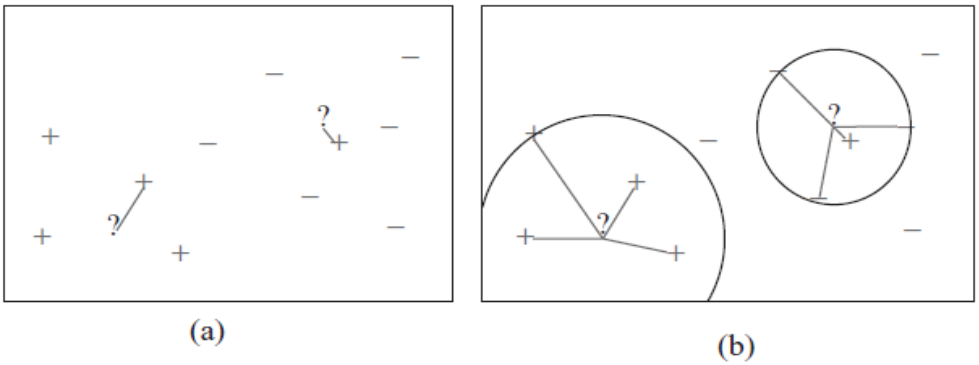


**Figura 1.5.** Clasificador kNN. Imagen recuperada de Mehmed Kantardzic.

### 1.5.3 Metodologías

El proceso de Minería de Datos aplicado en el proyecto académico de fin de carrera estará basado en la metodología propuesta por Mehmed Kantardzic [MEH 2011]. La importancia de esta metodología radica en tratar a la Minería de Datos como un proceso iterativo, en el cual se utilizarán diferentes técnicas y algoritmos para enriquecer los resultados del modelo. Además, los pasos establecidos están alineados con los objetivos del proyecto, los cuales incluyen integración, limpieza de datos y aplicación de los algoritmos. Por tal motivo, la metodología permitirá servir como referencia del desarrollo de Minería de Datos durante la realización del proyecto. A continuación, se mencionan las etapas que se planea seguir:

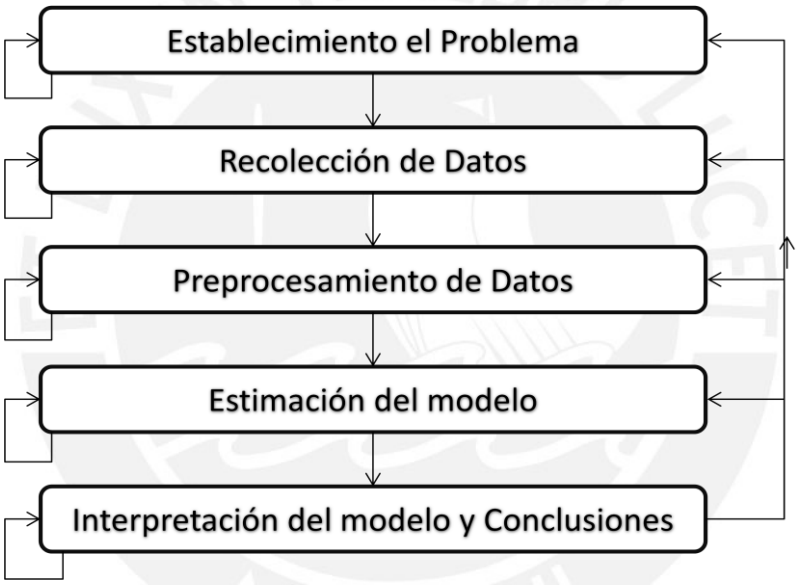


**Figura 1.6.** Proceso de Minería de Datos. Imagen adaptada de Mehmed Kantardzic.

#### 1) Establecimiento del Problema

El conocimiento y entendimiento del dominio en el que se va a realizar el proceso de Minería de Datos es de vital importancia, incluso aún más que la misma técnica de Minería de Datos, debido a que permite aclarar el problema que se desea tratar y determinar el rumbo al cual se tiene que enfocar el proyecto. En el proyecto académico de fin de carrera, esto se puede ver reflejado en la problemática y el estado del arte, donde se ha realizado el análisis de la situación actual de la

Minería de Datos en el sector salud público y como esto ha sido tratado en otros lugares del mundo [MEH 2011].

**2) Recolección de Datos**

Esta etapa está enfocada en la recolección de los datos desde las diversas fuentes que puedan existir en la entidad de salud y que sean necesarias para el aprendizaje del modelo. Para el caso del proyecto, se utilizará un alcance observacional para obtener la información; es decir, los datos que se recolectarán serán aleatorios de acuerdo con las condiciones del entorno. Sin embargo, se puede tener un conocimiento previo de dichas condiciones que permitan tener una idea de cómo estructurar el modelo e interpretar los resultados. Por esta razón, es importante tener el apoyo de los miembros de la entidad de salud que puedan colaborar con su experiencia para enfocarse en los datos verdaderamente importantes para el proceso de Minería de Datos [MEH 2011].

**3) Preprocesamiento de los Datos**

Una vez que los datos son recolectados desde alguna fuente de datos, pueden surgir una serie de inconvenientes que podrían afectar en los resultados del modelo. Por esta razón, es necesaria la implementación de técnicas que puedan lidiar con estos problemas en algunas ocasiones. Para el caso del proyecto, se realizarán las siguientes actividades que se encuentren alineadas con los objetivos [MEH 2011]:

- Creación de un DataMart de Pacientes: Permitirá integrar información del paciente que sea necesaria para el modelo y facilite su implementación.
- Detección y Eliminación de Outliers: Permitirá eliminar información que pueda degradar los resultados del modelo.
- Transformación de Datos Crudos: Incluye procesos de normalización, tratamiento de valores nulos, codificación, selección de características, etc. Esto se realiza con el fin de adecuar los datos originales a lo que necesitan las técnicas de Minería de Datos.

La etapa de Preprocesamiento debe estar vinculada con las demás etapas debido a que se trata de un proceso iterativo, donde se puede definir mejores conjuntos de datos para cada técnica de Minería de Datos en cada iteración. Está

comprobado que un buen método de Preprocesamiento brindará mejores resultados al modelo [MEH 2011].

**4) Estimación del Modelo**

En esta etapa se realiza la implementación de las técnicas de Minería de Datos. Para esto es muy importante tener en cuenta que exista coherencia entre el conjunto de datos y los algoritmos que se utilizarán. Como se ha mencionado anteriormente, el proyecto contempla el uso de varios algoritmos considerados buenos, para elegir finalmente el que otorgue mejores resultados; en este punto es donde se evidencia el proceso iterativo de la metodología de Minería de Datos escogida [MEH 2011].

**5) Interpretación del Modelo y Conclusiones**

Finalmente, la última etapa contempla interpretar los resultados del modelo y plasmarlos de tal forma que puedan ser entendidos por los encargados de tomar decisiones, en el caso del proyecto, los administradores de la entidad de salud. En este punto, es importante resumir los resultados con el fin que puedan contener solo información necesaria [MEH 2011].

## 1.6 Alcance

El proyecto de fin de carrera estará delimitado a predecir la continuidad del tratamiento de quimioterapia para pacientes que padecen de algún tipo de cáncer. Por esta razón, se ha pedido la colaboración del Instituto Nacional de Enfermedades Neoplásicas (INEN), el cual brindará las facilidades para poder recolectar la información histórica de los pacientes que servirán para el aprendizaje del modelo.

El abandono de los tratamientos es un problema frecuente en enfermedades complicadas, como lo es el cáncer, por lo cual se requiere de acciones que puedan mitigarlo [MIN 2013]; sin embargo, las estrategias no pueden ser enfocadas de la mejor manera debido a que no se cuenta con el conocimiento de que pacientes tienen una mayor tendencia a abandonar su tratamiento. Por tal motivo, el producto final del proyecto de Minería de Datos pretende servir de apoyo a la toma de decisiones de los administradores del hospital y personas a cargo de las estrategias. Como elemento

principal del proyecto, la herramienta permitirá realizar las predicciones periódicamente de forma automatizada. Para esto, se utilizarán los algoritmos de clasificación ya implementados de la librería Weka.

Finalmente, el proyecto de fin de carrera contemplará las etapas mencionadas en la metodología de forma que puedan ser automatizadas y adaptables a futuro, con el fin de que la herramienta pueda ser útil y precisa a largo plazo.

### 1.6.1 Limitaciones

Entre las principales limitaciones que podrían dificultar el desarrollo del proyecto se encuentran:

- Falta de colaboración del área de Sistemas de la entidad de salud para proveer información acerca del diseño actual de la Base de Datos, e incluso, para poder integrar el proyecto dentro de la organización. Esto podría detener el proyecto en su etapa primordial, como es la recolección de datos.

- La información almacenada en los repositorios es bastante grande y no se encuentra adecuadamente organizada para la explotación de Minería de Datos. Esta limitación podría extender los tiempos de desarrollo de la etapa de pre-procesamiento, dependiendo de qué tan complejo sea el análisis. Por esta razón, se requeriría un proceso de integración de datos, con la información útil para el modelo.

### 1.6.2 Riesgos

A continuación, se presentan los riesgos del proyecto y los planes de acción antes y después de que estos ocurran.

**Tabla 1.2**. Cuadro de Riesgos del proyecto. Elaboración propia.

| N° | Riesgo | Probabilidad | Impacto | Severidad | Descripción | Planes de Mitigación | Planes de Contingencia |
|---|---|---|---|---|---|---|---|
| 1 | El software a utilizar sea muy complejo para aprender | MEDIA | ALTO | ALTO | Las herramientas de Software demandan una gran cantidad de tiempo y habilidad para poder aprender a usarlas | Elegir herramientas de Software cuyo aprendizaje no sea tan complejo o que el tesista ya tenga conocimiento acerca de ellas | Identificar los elementos del Software que se utilizarán principalmente en el proyecto y enfocarse en aprender a usar al menos esos. |
| 2 | Los miembro de la entidad pública se encuentren en huelga médica | MEDIA | MEDIO | MEDIO | Los miembros de la entidad pública no se encuentran disponibles para permitir el acceso al instituto | Identificar a los miembros que apoyan el proyecto y obtener los teléfonos y correos para contactar en caso exista un imprevisto | Mantener contacto con los miembros de la entidad pública para coordinar reuniones que no se vean afectadas por huelgas en el sector salud |
| 3 | Falta de tiempo para la presentación de los entregables | MEDIA | ALTO | ALTO | El tesista está sobrecargado de tareas y no tiene tiempo para culminar los entregables en las fechas pactadas | Realizar una planificación del desarrollo del proyecto para cada entregable, que pueda cumplir con los requisitos del proyecto de fin de carrera. Además, se debe avanzar el proyecto durante Julio y Agosto | Terminar el entregable pasado e integrarlo con el nuevo entregable. Si falta demasiado tiempo se sugiere retirarse del curso de proyecto de Tesis |
| 4 | El área de sistemas de la institución no colabora con brindar la información para el modelo | MEDIA | ALTO | ALTO | Los miembros encargados de facilitar la interacción con la Base de Datos no otorgan las facilidades para la etapa de recolección de datos. | Enviar una solicitud a la entidad de salud para que permita la autorización y desarrollo del proyecto de fin de carrera | Conversar con los miembros del área de Sistemas para llegar a un acuerdo, conforme a la autorización brindada por la entidad |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | La base de datos se encuentra desorganizada y con muchas inconsistencias. | ALTA | MEDIO | ALTO | La Base de Datos presenta dificultad para comprender su diseño y contiene datos que pueden distorsionar el modelo | Solicitar información al área de sistemas acerca del diseño de la Base de Datos e información relacionada | Desarrollar un procedimiento de organización de la información que necesitamos dentro del proceso de limpieza de datos |
| 6 | El asesor no se encuentra disponible | BAJA | MEDIO | MEDIO | El asesor no tiene el tiempo adecuado para responder a las consultas del tesista | Consultar acerca de la disponibilidad de los asesores relacionados con el proyecto | Cambiar a un asesor que se encuentre disponible |
| 7 | Los algoritmos utilizados no proveen los resultados esperados | MEDIA | ALTO | ALTO | Los algoritmos que se decidieron usar otorgan resultados de precisión por debajo de lo esperado | Realizar una investigación profunda acerca de los algoritmos usados en problemáticas similares y que puedan cumplir con los requerimientos del proyecto | Implementar algoritmos diferentes, de acuerdo con investigaciones pasadas, y realizar la etapa de preprocesamiento de forma más exhaustiva |
| 8 | La laptop donde se encuentra el proyecto sea robada | BAJA | ALTO | MEDIO | El dispositivo donde el tesista almacena los archivos del proyecto (ejecutables, documentos, bibliografía, etc) es robado | Utilizar medios de almacenamiento en un repositorio en nube o usb, con el fin de tener un respaldo del proyecto | Buscar los últimos respaldos del proyecto que se tengan disponibles y los entregables que posea el asesor,y rehacerlo desde tal punto |
| 9 | Cambio en los alcances a mitad del proyecto | MEDIA | ALTO | ALTO | A mitad del proyecto, se decide modificar los requerimientos del proyecto | Plantear los requisitos del proyecto de manera que sean alcanzables durante el tiempo establecido | Modificar los puntos necesarios para que puedan adecuarse el nuevo alcance del proyecto. Si falta tiempo, se sugiere retirarse del curso |

## 1.7 Justificativa y Viabilidad del Proyecto

### 1.7.1 Justificativa

Diariamente, las instituciones de salud procesan grandes cantidades de información, las cuales, si son explotadas bajo los métodos correctos, podrían generar conocimiento de gran valor para los administradores de dicha institución. Una forma de aprovechar el contenido de la información es utilizar una técnica de Minería de Datos. La Minería de Datos se ha convertido en un elemento muy importante en el sector salud debido a que sirve como soporte en la gestión óptima de las entidades de salud.

En el contexto de la salud pública peruana, el uso de la Minería de Datos es un tema nuevo para muchas instituciones, por lo cual, existen varias necesidades, actualmente, que se pueden cubrir con el uso de estas técnicas. La necesidad o problema en el cual se enfoca el proyecto es el comportamiento del paciente con respecto al tratamiento. Se ha identificado que en enfermedades críticas como el cáncer, cuyo índice de ocurrencia ha amentado en los últimos años, la tasa de abandono del tratamiento es alta, lo cual podría generar mayor deterioro en la salud del paciente y problemas en la planificación de los recursos destinados para dichos tratamientos [MIN 2013]. Por esta razón, la importancia del proyecto radica en implementar un modelo predictivo que pueda determinar a los pacientes que presentan la tendencia de abandono.

Poder conocer a los pacientes que presentan tendencias de abandonar el tratamiento permitirá que se puedan tomar acciones de manera personalizada, motivándolos a que continúen sus tratamientos. Estos conocimientos que generará el modelo permitirán apoyar las decisiones de los administradores de la entidad de salud de manera más efectiva para el problema planteado.

Finalmente, la precisión del modelo predictivo también es vital para que este nuevo conocimiento pueda ser utilizado por la entidad de salud. Cada etapa planteada en el proyecto buscará lograr este objetivo. De acuerdo con Mehmed [MEH 2011], la etapa de recolección de datos confiables y el pre-procesamiento de estos son críticas para lograr buenos resultados, aún más que el mismo algoritmo; por tal motivo, el proyecto también buscará reforzar estas etapas a través de los métodos y procedimientos mencionados anteriormente. En el caso del resultado que plantea la creación de un DataMart, esto también se debe a que el INEN no cuenta con un DataWarehouse con

información integrada y organizada; por tal motivo, el DataMart podría servir como primer paso para una mejor estructuración de la información en un futuro.

### 1.7.2 Viabilidad

**Viabilidad Técnica**

Para el proyecto se utilizará la metodología de Minería de Datos basada en el libro de Mehmed Kantardzic, para lo cual se ha planteado una serie de etapas alineadas con los objetivos del proyecto. Cada etapa requiere de los métodos y procedimientos asignados anteriormente, varios de los cuales son conocidos por el tesista y son apoyados por varias herramientas de Minería de Datos.

Además, de acuerdo con las investigaciones revisadas en el estado del arte, se ha podido determinar que el desarrollo de la Minería de Datos es factible dentro del contexto de la salud pública. Por tal motivo, se puede planificar adecuadamente el proceso que tendrá el proyecto, siempre y cuando se apliquen las recomendaciones de la metodología y las buenas prácticas encontradas en los estudios realizados. Cabe destacar que se necesita la participación de los miembros de la institución, por lo cual se ha realizado la solicitud respectiva para poder desarrollar el proyecto, la cual ha sido aprobada.

Con respecto a las herramientas, se ha definido un grupo de Software de código abierto cuyo aprendizaje no será excesivamente complicado debido a que existe bastante documentación sobre su uso y, además, han sido utilizados en varios proyectos con buenos resultados.

En conclusión, los métodos, procedimientos y herramientas apoyarán el cumplimiento de cada resultado esperado en el proyecto, por lo cual, se puede determinar que será técnicamente viable.

**Viabilidad Temporal**

El presente proyecto tendrá, aproximadamente, una duración total de 6 meses. Cabe mencionar que durante los primeros 4 meses se realizará la implementación del modelo predictivo completo (lo cual permitirá que los objetivos sean logrados por el

tesista en el tiempo establecido por el curso Proyecto de Tesis II), dejando los últimos 2 meses para que la efectividad del modelo pueda ser probada con los datos de nuevos pacientes en la entidad de salud pública. La mayor parte del tiempo se requerirá de coordinación con el área de Sistemas para acceder a las fuentes de información y el pre-procesamiento dentro de la misma institución; por tal razón, se vienen acordando reuniones con los miembros involucrados en el proyecto para acordar estos temas importantes de manera planificada.

**Viabilidad Económica**

Con respecto a las necesidades económicas para el proyecto, se puede mencionar el uso de materiales, movilidad y recursos humanos.

Los gastos en materiales corresponderán a la documentación realizada para cada iteración del proyecto. Además, se requerirá de una laptop personal, computadoras dentro de la entidad de salud, servicio de red, servicio de electricidad, entre otros. Los gastos en movilidad ocurrirán cuando exista la necesidad de trabajar dentro de la entidad, ya que dicha información debe permanecer confidencial y dentro de la institución, o cuando existan reuniones con los miembros involucrados en el proyecto. Con respecto a los recursos humanos, estos corresponden al tiempo que invertirá el tesista en la realización del proyecto, lo cual representa un gasto menor para la institución en comparación a adquirir una solución de Software especializada.

Para terminar, como se menciono anteriormente, las herramientas de Minería de Datos que serán utilizadas son de licencia libre o “Open Source”, lo cual significa que por el uso de estas no se generarán gastos. Esto significa que, económicamente, el proyecto será viable tanto para el tesista como para la entidad de salud pública.

# CAPÍTULO 2

## 2.1 Marco conceptual

La Minería de datos, siendo un proceso que tiene un conjunto de etapas dependientes unas de otras, presenta conceptos complejos para cada una de las etapas, los cuales están relacionados con los problemas específicos que se pretende tratar. Estos conceptos deben relacionarse con el contexto que se pretende analizar, el cual será el sector salud.

### 2.1.1 Objetivo del marco conceptual

Esta sección tiene como objetivo definir los conceptos relevantes para el entendimiento de la problemática del proyecto de fin de carrera. Se ha optado por separar los conceptos relacionados con el proceso de Minería de Datos y los procesos propios del sector salud.

### 2.1.2 Conceptos relacionados con Minería de Datos

**Minería de Datos**

La Minería de Datos es el proceso que recibe como entrada un conjunto de datos y lo convierte en conocimiento. El proceso se basa en encontrar patrones de comportamiento útiles en los datos [WEI 2010].

Existen dos concepciones erradas acerca de la Minería de Datos: 1) consiste en grandes cantidades de información aisladas entre sí esperando que ocurra un problema; 2) El problema está ligado a una sola técnica o algoritmo. Por el contrario, la Minería de Datos se enfoca en la integración de los datos para adquirir un mayor conocimiento del ambiente donde se desarrolla el problema. Además, se considera un proceso iterativo ya que permite el uso de varias técnicas hasta encontrar la que mejor se pueda adecuar al modelo [MEH 2011].

La Minería de Datos es una etapa perteneciente al descubrimiento de conocimiento en base de datos (KDD) que consiste en el análisis de los datos y uso de algoritmos para encontrar un conjunto de patrones para esos datos [FAY 1996].

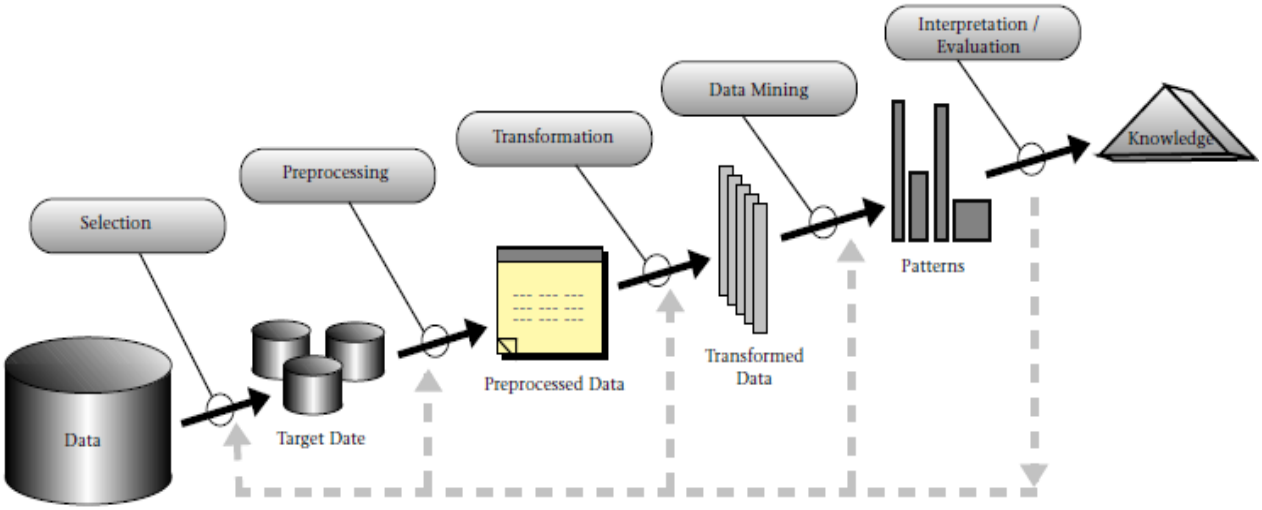


**Figura 2.1.** Pasos que constituyen el proceso de descubrimiento de conocimiento. Imagen recuperada de Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth (1996).

El proceso general descrito por Fayyad, Piatetsky- Shapiro y Smyth consta de 5 etapas [FAY 1996]:

- El proceso de selección indica como la data es generada y recolectada. Un conocimiento a priori acerca de cómo los datos están distribuidos y relacionados nos podría dar las pautas para escoger la información necesaria para el modelo.
- Los datos, generalmente, vienen crudos de algún Datamart o Data warehouse. El pre procesamiento se encarga de identificar las variables necesarias para y descartar datos que no sean útiles o tengan una influencia negativa en el modelo.
- Una vez que se tienen los datos necesarios, estos pasan por el proceso de transformación, el cuál modificará las muestras para que puedan adecuarse al modelo.
- La cuarta etapa consiste en la Minería de Datos. En esta etapa se utiliza el algoritmo para determinar patrones en los datos
- Finalmente, los resultados del algoritmo nos proveerán el conocimiento acerca del problema para tomar decisiones acertadas.

Los algoritmos de Minería de Datos son, generalmente, escalables; es decir, pueden adecuarse y tener un buen rendimiento sin limitaciones en el conjunto de datos. Muchos de estos algoritmos fueron diseñados con la idea de trabajar con valores de todo tipo [WEI 2010].

Las principales tareas en que se enfoca la Minería de Datos son las siguientes:

a) Maquinas de Aprendizaje: Muchos de las tareas de aprendizaje han sido imitadas de sistemas biológicos como el lidiar con un ambiente desconocido, propio del desarrollo de los humanos. Las Maquinas de aprendizaje utilizan éstos métodos para aprender de un ambiente desconocido (inducción), y en base a ello, estimar un posible resultado para nuevas entradas (deducción). El concepto se tratará más a fondo dentro de este marco conceptual [MEH 2011].
b) Web Mining: Es el proceso de descubrir y extraer información automáticamente de documentos y servicios de Internet. La complejidad de esta tarea radica en la falta de una estructura definida en los contenidos de Internet y a la gran cantidad de páginas que son creadas diariamente [MEH 2011].
c) Text Mining: Muchas organizaciones tienen almacenados grandes cantidades de información en documentos, algunas veces sin usar. Text Mining se enfoca en el descubrimiento de nueva información para una colección de documentos de texto [MEH 2011].

**Machine Learning (Máquina de aprendizaje)**

Una Máquina de aprendizaje se puede resumir como el aprendizaje del comportamiento de una cantidad de variables de un conjunto de datos para formar un patrón general que pueda representarlas. La mayor parte de las tareas de las máquinas de aprendizaje se basan en el aprendizaje inductivo [MEH 2011]. El proceso básico de aprendizaje requiere tres componentes:

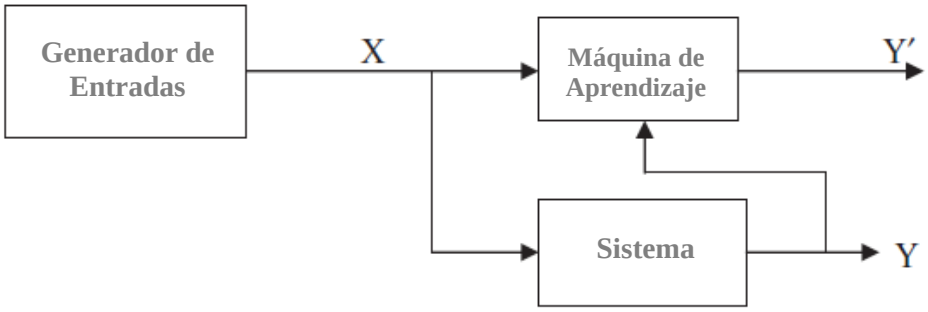


**Figura 2.2.** Máquina de aprendizaje. Imagen adaptada de Mehmed Kantardzic.

- Un conjunto de vectores de entrada aleatorios X: La máquina de aprendizaje no recibe un esquema de muestras definidos, sino un conjunto de valores de entrada desconocidos que serán provistos al sistema. Por esta razón, se puede considerar que el aprendizaje inductivo es de carácter observacional [MEH 2011].
- Un sistema que, dado un vector de entrada X, devuelva una salida Y: El sistema actuará como si fuera exacto, por lo que devolverá un vector Y que puede ser representado como un F(X); sin embargo, puede existir el caso que existan algunas entradas no medidas, las cuales pueden ser representadas como valores aleatorios utilizando una variable de probabilidad con la función F(X). Tanto X e Y servirán para alimentar a la máquina de aprendizaje [MEH 2011].
- Una máquina de aprendizaje: El objetivo principal es estimar una salida Y' basándose en la información provista por el sistema. Para lograr esto, se realiza una generalización del comportamiento del sistema (conjunto de datos de entrenamiento) utilizando un conocimiento a priori. Como resultado, la máquina de aprendizaje devolverá una función H(X) o Y' que se pueda aproximar al conjunto de funciones F(X) [MEH 2011].

De acuerdo con [MEH 2011], el aprendizaje inductivo se puede clasificar en dos tipos: aprendizaje supervisado y aprendizaje no supervisado.

El aprendizaje supervisado consiste en encontrar una relación o patrón de un conjunto de entrenamiento con un X (entrada) y F(X) (salida) conocidos. Por lo tanto, se tendrá el conocimiento del entorno que se desea analizar. Cuando un X' con un F(X') desconocido entre al modelo, el aprendizaje supervisado permitirá estimar el valor de F(X') basándose en el conocimiento adquirido. Un conjunto de pruebas, con un F(X')

conocido pero que entra el modelo como desconocido, es utilizado para calcular la precisión de este tipo de aprendizaje. La diferencia entre el F(X') estimado y el F(X') es conocido como señal de error y permitirá determinar si el modelo es bueno. Las principales tareas que se basan en el aprendizaje supervisado son las siguientes [MEH 2011]:

a) **Clasificación:** Permite, para un conjunto de clases definidas, predecir a cuál de ellas pertenece una muestra. La complejidad de obtener una buena solución dependerá de la cantidad de dimensiones y la diferenciación entre los comportamientos de cada clase [MEH 2011]. El conjunto de entrenamiento provee dos tipos de información: el vector X, el cual es el comportamiento de la muestra y F(X), el "atributo objetivo", que representa la clase a la que pertenece X. Un problema de clasificación puede ser "binario", cuando se desean respuestas como si o no, y "multiclase", cuando existen más de dos categorías. Entre los algoritmos que permiten desarrollar el proceso de clasificación están: arboles de decisión, bayesiano, SVM y redes neuronales [HOR 2007].

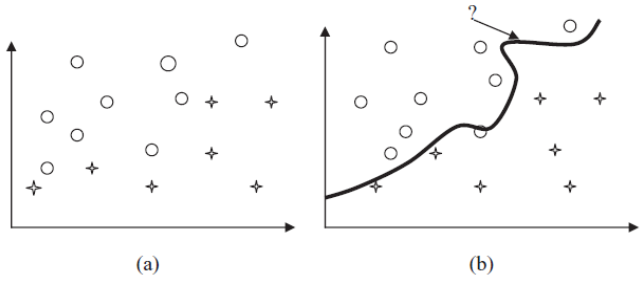


**Figura 2.3.** Interpretación gráfica de Clasificación. Imagen recuperada de Mehmed Kantardzic.

b) **Regresión:** Está enfocado en la predicción de valores numéricos continuos. A diferencia de la clasificación, el "atributo objetivo" F(X) representa un valor numérico real que puede ser diferente al de las demás muestras. Por ejemplo, una de sus aplicaciones es calcular el precio, como el de una casa, bajo ciertas características, como el número de habitaciones o baños. Los algoritmos que permiten implementar este tipo de aprendizaje son: árboles de decisión, redes neuronales, SVM, regresión lineal y modelos lineales generalizados (GLM) [HOR 2007].

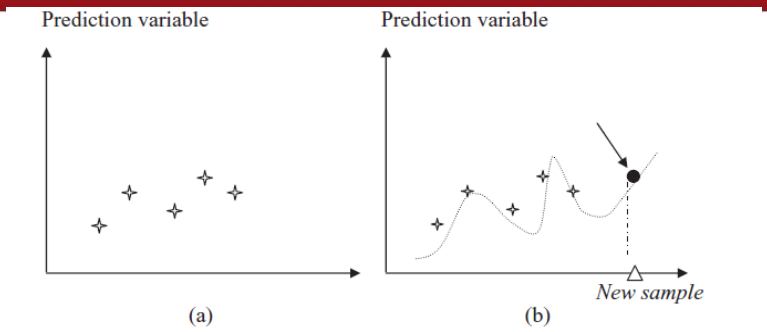


**Figura 2.4.** Interpretación gráfica de Regresión. Imagen recuperada de Mehmed Kantardzic.

El aprendizaje no supervisado consiste en descubrir patrones para un único conjunto de entrada X; en este caso, no se tiene conocimiento de F(X). El sistema utiliza la entrada para aprender acerca del ambiente, y luego genera representaciones globales o locales del conjunto de datos [MEH 2011]. Las principales tareas que utilizan el aprendizaje no supervisado son:

**a) Clustering:** Está enfocado en identificar grupos en un conjunto de datos que presentan comportamientos similares entre sus respectivas muestras. Como resultado, obtendremos la variable de salida F(X), la cual representa el grupo al que pertenece la muestra. El proceso de clustering será bueno si es que existe una alta diferencia entre los grupos formados. Entre sus aplicaciones destacan la segmentación de clientes, detección de outliers, text mining, entre otros. Entre los algoritmos que apoyan el clustering se encuentran: k-means, mapas auto organizados, clustering de partición ortogonal y clustering jerárquico [HOR 2007].

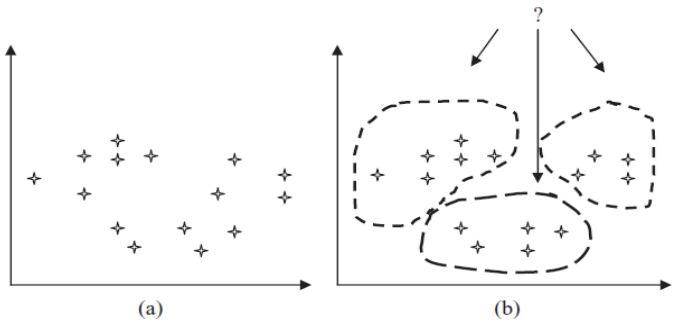


**Figura 2.5.** Interpretación gráfica de Clustering. Imagen recuperada de Mehmed Kantardzic.

**b) Asociación:** Está enfocado en descubrir relaciones, representadas como reglas, entre los elementos de un conjunto de datos [WEI 2010]. Está basado en una relación de causa-efecto, donde si un elemento A existe, otro elemento B también debe existir. Entre sus aplicaciones más usadas está el análisis de

mercados (market basket); por ejemplo, determinar relaciones entre la compra de productos como leche y pañales [HOR 2007].

| Association Rule | Support | Confidence |
|---|---|---|
| {Ketchup} → {Soda} | 0.4 | 1.0 |
| {Cereal} → {Milk} | 0.4 | 1.0 |
| {Greeting Card} → {Cake} | 0.4 | 1.0 |
| {Cake} → {Greeting Card} | 0.4 | 1.0 |

**Figura 2.6.** Asociación. Imagen recuperada de Gary Weiss.

c) **Summarization:** Está enfocado en encontrar límites para conjuntos o subconjuntos de datos. Su utilidad radica en que permite simplificar y mejorar la toma de decisiones en el dominio que se realizó el proceso [MEH 2011].

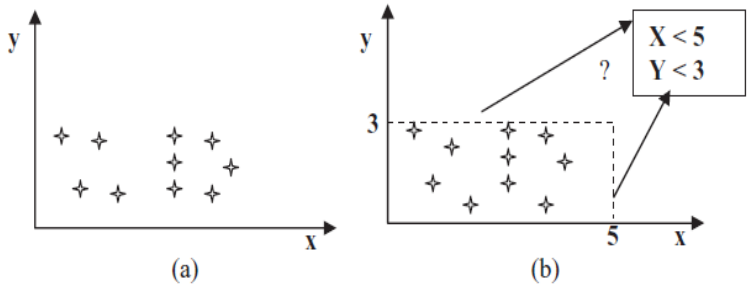


**Figura 2.7.** Interpretación gráfica de Summarization. Imagen recuperada de Mehmed Kantardzic.

d) **Modelo de dependencia:** Está enfocado en descubrir modelos para subconjuntos de datos dentro del conjunto de entrenamiento. Permite identificar dependencias entre características o entre muestras. Su utilidad es apreciada cuando se trabaja con conjuntos de datos grandes ya que reduce la complejidad computacional al trabajar con subconjuntos específicos [MEH 2011].

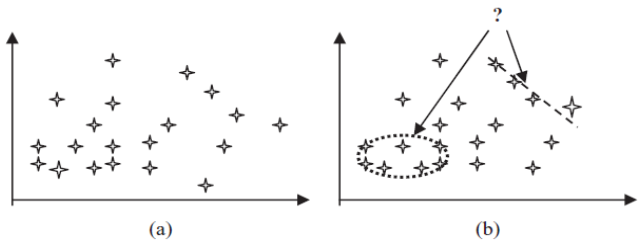


**Figura 2.8.** Interpretación gráfica de Modelo de dependencia. Imagen recuperada de Mehmed Kantardzic.

e) **Detección de cambio y desviación:** Está enfocado en la identificación de cambios significativos en un conjunto de datos. Por ejemplo, la detección de outliers [MEH 2011].

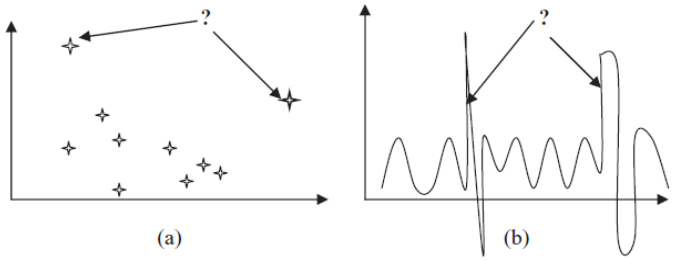


**Figura 2.9.** Interpretación gráfica de Detección de cambio y desviación. Imagen recuperada de Mehmed Kantardzic.

### 2.1.3 Conceptos relacionados con el sector salud

**Diagnóstico**

De acuerdo con la Real Academia Española [RAE 2014], Diagnóstico es el proceso que permite identificar el carácter de una enfermedad a través de evaluaciones. Cada vez que un paciente asista a una consulta, se realizará un "acto médico", lo cual comprende las acciones de diagnóstico, terapéuticas y de pronóstico de la enfermedad que el paciente padece [MIN 2005].

De acuerdo a la norma técnica de la historia clínica de los establecimientos de salud [MIN 2005], el diagnóstico puede ser de los siguientes tipos:

- Dependiendo de su confirmación:
  - Presuntivo: Requiere algunos exámenes y evaluaciones adicionales para confirmar la enfermedad que el paciente tiene.
  - Definitivo: Se confirma la enfermedad y se puede iniciar el tratamiento correspondiente
- Diagnóstico producto de una atención médica
- Diagnóstico de discapacidad
- Otros diagnósticos como nutricional, de salud mental, de riesgos, etc.

Los diagnósticos, después de cada consulta, deberían seguir el CIE-10. El CIE-10 o "Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados

con Salud-Décima Revisión" representa una clasificación estándar para todos los problemas epidemiológicos y temas concernientes a la gestión de la salud [PAH 2003].

**Tratamiento**

Según la RAE [RAE 2014], Tratamiento es el proceso que emplea un conjunto de técnicas destinadas a aliviar o curar una enfermedad. Un tratamiento debe poseer las siguientes características [MIN 2005]:

- El tratamiento debe ser claro.
- Se deben especificar todas las indicaciones terapéuticas, lo cual incluye cuidados, alimentación, medicamentos, dosis, entre otros.
- Se debe especificar la fecha de cada consulta.
- Se debe anotar la evolución del tratamiento.

El tratamiento empieza desde el primer diagnóstico para una enfermedad determinada hasta que el paciente deja de asistir a sus citas y lo abandona. Se considera como abandono del tratamiento cuando el paciente no acude a su siguiente consulta durante un periodo de tiempo establecido para cada tipo de enfermedad [UNF 2012]. Los motivos pueden ser los siguientes:

- Cura de la enfermedad (el paciente fue dado de alta por el médico).
- Fallecimiento
- Abandono voluntario del tratamiento

**Factores principales en el abandono del tratamiento**

A continuación se describen algunos factores que se prevé serán influyentes en el comportamiento de los pacientes durante su tratamiento. Dichos factores fueron recopilados de algunos estudios realizados a pacientes que asistían a terapia en Australia [ROB 2013]. Si bien estos factores serán influyentes, no serán los únicos que participarán en el proceso de Minería de Datos; esto será determinado por el análisis de componentes principales (PCA) definido previamente.

- Indicadores Socioeconómicos: Consiste en las variables que dependen del contexto social y económico; por ejemplo, economía, educación, entre otros.
- Indicadores Demográficos: Consiste en las variables que dependen de la población; por ejemplo, edad, sexo, residencia, entre otros.
- Tipo de enfermedad: Define los tipos y sub tipos de cada enfermedad.
- Grado de enfermedad: Define la etapa en la cual se encuentra la enfermedad. Esta información será brindada a través de un diagnostico.

### 2.1.4 Conclusión

Los conceptos definidos previamente servirán de apoyo para entender la problemática que abordará el proyecto de fin de carrera. Los conceptos propios de la Minería de Datos se podrán apreciar durante el proceso de diseño e implementación. Los conceptos relacionados con el sector salud servirán para entender el problema central que se pretender tratar y los resultados que se obtengan al final del proceso de Minería de Datos.

## 2.2 Estado del arte

En los últimos años, la Minería de Datos ha desarrollado un papel muy importante en el sector salud, por lo cual muchas instituciones de salud han optado por utilizar técnicas de Minería de Datos para resolver sus principales problemas [HOR 2007]. La gran diversidad de tareas que puede desarrollar la Minería de Datos permite su uso en distintos problemas y enfoques. Esto convierte a la Minería de Datos en un proceso eficiente y flexible en el sector salud.

La gran cantidad de investigaciones desarrolladas en el sector salud han apoyado en la planificación de los recursos del hospital y han servido como indicador para brindar un mejor servicio a los pacientes.

### 2.2.1 Objetivos de la revisión del estado del arte

El objetivo del Estado del Arte es identificar investigaciones pasadas que aborden, de forma similar, el problema del comportamiento de los pacientes durante su tratamiento en una entidad de salud pública. Las investigaciones encontradas permitirán conocer como ha sido tratado el tema y servirán de apoyo para tener una idea clara de cómo se desarrollará el problema en el proyecto de fin de carrera.

### 2.2.2 Método usado en la revisión del estado del arte

El método utilizado en la revisión del estado del arte fue la revisión tradicional.

### 2.2.3 Uso de Técnicas de Minería de Datos para Determinar y Predecir el Periodo de Estancia de Pacientes Cardiacos

El estudio titulado “Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients” fue realizado por doctores de la Universidad de Ciencias Médicas de Tehran en Irán [HAC 2013]. El objetivo principal fue implementar una técnica de Minería de Datos capaz de predecir con alta precisión el tiempo de duración de los tratamientos para pacientes con problemas cardiacos. Esto con el fin de brindar a los pacientes mejores servicios y satisfacción durante el proceso de su hospitalización.

Una característica identificada en los pacientes con problemas cardiacos es que permanecen por largos periodos de tiempo hospitalizados o recibiendo un tratamiento. Esto genera que se utilicen muchos recursos del hospital, lo cual conduce a un exceso de costos si es que no se tiene en cuenta una buena planificación para distribuirlos en todos los pacientes. Por esta razón, una forma de evaluar el tiempo en que un paciente estaría hospitalizado, dependiendo del tipo de enfermedad que padecía, era necesaria para manejar eficientemente la asignación de sus cuartos de descanso y para que los administradores del hospital puedan tomar mejores decisiones.

El estudio tuvo la participación de 4,948 pacientes con problemas cardiacos (enfermedad de la arteria coronaria). Se escogieron unas 36 variables de entrada para el modelo predictivo basándose en la situación actual de los pacientes y estudios

pasados que trataban, con métodos estadísticos, los principales factores de la mortalidad hospitalaria. El conjunto de datos final contó con los registros de 2,065 pacientes debido a que sólo se contó los que tenían la enfermedad de arteria coronaria y se realizaron procesos de limpieza de datos, eliminación de registros con variables nulas y detección de valores anómalos. Finalmente, se separó un 80% del conjunto de datos (1,643 muestras) para el conjunto de entrenamiento y un 20% (421 muestras) para el conjunto de pruebas.

El proceso de Minería de Datos fue totalmente iterativo debido a que participaron cuatro algoritmos con el mismo conjunto de datos para obtener el que proporcionará predicciones con mayor precisión. Los algoritmos fueron:

1) Redes Neuronales Artificiales: Utilizado para encontrar patrones lineales y no líneas dentro del conjunto de datos. Es uno de los algoritmos que posee mejor rendimiento en las aplicaciones de medicina.
2) Support Vectors Machine (SVM's): Es un algoritmo muy potente que permite clasificar conjuntos de datos lineales y no lineales gracias a su criterio de optimización. Dependiendo de la configuración de sus parámetros, se puede obtener los factores con mayor influencia en el modelo.
3) Árbol de decisión: Representa un árbol de conocimiento, donde en las hojas se encuentran las reglas de clasificación en las ramas la salida o resultado. Son considerados como algoritmos confiables y útiles dentro de la toma de decisiones clínica ya que permiten obtener fácilmente las reglas de clasificación (causa y efecto), como se explicó en la tarea de Minería de Datos conocida como asociación.
4) Modelos Combinados: Utiliza la combinación de los tres algoritmos con el fin de aprovechar los beneficios de cada uno.

Los resultados otorgaron al algoritmo SVM un porcentaje de acierto de 96.4%, siendo el de mayor precisión), al algoritmo de modelos combinados un 95.9%, al algoritmo de árbol de decisión un 83.5%, y al algoritmo de redes neuronales un 53.9%. Además, gracias al algoritmo SVM, se pudo determinar los factores más influyentes y, gracias al algoritmo de árbol de decisión, las reglas asociadas a las categorías de tiempo de duración de la hospitalización.

En resumen, tres de los cuatro algoritmos obtuvieron una alta precisión al momento de predecir; por lo tanto, se puede afirmar que el modelo es confiable para determinar el

tiempo de duración de la hospitalización. Además, el algoritmo SVM, el cual fue el que dio mejor resultado, tiene la característica de poder brindar información acerca de los factores más influyentes en la elección tomada por el modelo. Este conocimiento será de gran apoyo para la toma de decisiones en el hospital y servirá para ofrecer un servicio personalizado a los pacientes.

### 2.2.4 Factores de Deserción de un Programa Pediátrico de Control de Peso

La división de Psicología y Cardiología de la Universidad de Medicina de Cincinnati en Ohio realizó un estudio titulado “Predictors of attrition from a pediatric weight management program” que tenía como objetivo determinar los principales factores que llevan a que una familia decida abandonar el tratamiento de obesidad de su hijo para aplicarlo con nuevos pacientes que pueden ser niños o adolescentes [ZEL 2004].

Las muestras recolectadas fueron de 212 pacientes, entre niños y adolescentes, que tenían un índice de masa muscular elevado. Estos pacientes fueron sometidos a una serie de evaluaciones fisiológicas, nutricionales y psicológicas con el fin de obtener las variables que servirían de entrada para el modelo.

El proceso de Minería de Datos se realizó con un método estadístico, conocido como ANOVA, el cual permite identificar que entradas están relacionadas con una salida predeterminada [MEH 2011]; para este caso, que factores tienen mayor relación con el hecho de no continuar en un tratamiento. Debido a que se utilizó este método estadístico, una transformación de los datos no numéricos tuvo que ser necesaria para que sean aceptados por el modelo.

El resultado del estudio dio a conocer que el 55% de los pacientes abandonaba su tratamiento. Adicionalmente, se pudo determinar los factores más influyentes en la decisión de estos pacientes, como son la edad, la autoestima, los medicamentos suministrados, entre otros.

Para terminar, se concluye que los resultados serán de gran utilidad para tratamientos de obesidad futuros ya que podrán ser abordados de acuerdo a las características que se mostraron como influyentes en el estudio. Los pacientes con tendencia a abandonar sus tratamientos podrían recibir un apoyo personal, junto con una estrategia de retención.

### 2.2.5 Uso de Minería de Datos para Describir el Tiempo de Estancia en un Hospital

El estudio “Using Data Mining to Describe Long Hospital Stays” fue realizado en Colombia por miembros de la Universidad de los Andes [GOM 2009]. El estudio tiene como objetivo utilizar una técnica de Minería de Datos que permita predecir el tiempo de estancia en el hospital y el entendimiento de todo el proceso de Minería de Datos que se lleva a cabo.

Las muestras fueron recogidas en el 2006 en archivos de texto donde se podía obtener datos de consultas, diagnósticos, cirugías y tiempos de estancia, la cual permite clasificar a cada paciente de acuerdo a nuestro objetivo. Dichos datos pasaron por un proceso de integración debido a las diferentes fuentes de datos que existían. Además, se realizó una normalización con el método de búsqueda aproximada, cuyo fin era aumentar la uniformidad en los formatos de los datos para hacer más eficiente el uso de los algoritmos escogidos. El algoritmo utilizado fue reglas de asociación debido a que se adaptaba a las condiciones del modelo y requerimientos del resultado. Esto era importante para determinar qué factores eran más influyentes en el tiempo de estancia ya que el algoritmo permite encontrar las relaciones entre las variables.

Como resultado, se pudo obtener factores influyentes que ya eran conocidos, como el nivel de medicamentos, y otros que no hubiera sido posible determinar de manera trivial; por ejemplo, la influencia del sexo del paciente.

En resumen, los resultados obtenidos permitirán modificar los criterios del límite de estancia en el hospital de acuerdo al grupo de pacientes que se llegue a hospitalizar. Esto permitirá distribuir de forma eficaz los recursos del hospital, diferenciando los casos de hospitalización simples de los complejos, y brindar un mejor servicio a cada paciente.

### 2.2.6 Aplicación de Máquina de Vector de Soporte para la Predicción del Cumplimiento de la Medicación en Pacientes con Insuficiencia Cardiaca

Soo-Kyoung Lee, miembro de la Universidad Nacional de Seúl en Corea del Sur, realizó una investigación, cuyo título era “Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients”, en pacientes con

insuficiencias cardiacas con el fin de identificar los factores críticos que afectaran el compromiso de dichos pacientes con el tratamiento de medicación [SON 2010]. Para esta investigación se optó por emplear una máquina de aprendizaje: clasificador (tarea de Minería de Datos).

El conjunto inicial de datos fue de 96 pacientes. Luego de realizar una limpieza de datos y reducción de valores, 76 muestras de pacientes fueron las que se tomaron en cuenta para el modelo. Además, se tomaron en cuenta 11 variables de carácter clínico y demográfico. La clase objetivo se dividió en dos categorías: pacientes que se adhieren a la medicación y el resto, los que no.

El algoritmo que se utilizó fue el Support Vector Machine (SVM) debido a su alto poder de clasificación en escenarios donde la cantidad de muestras son pequeñas y existen un gran número de variables. En la elección del algoritmo, se tomaron en cuenta resultados de investigaciones pasadas donde el algoritmo SVM tuvo un alto rendimiento.

Las variables fueron divididas en varias combinaciones para determinar los grupos que presentaban una tasa de acierto más alta. Como resultado, se pudo obtener una precisión de 77.632%, lo cual es considerado como un grado de utilidad moderado para los fines de la institución de salud.

Finalmente, el conocimiento adquirido con el clasificador permitirá poder realizar acciones personalizadas con los pacientes que no cumplan con su proceso de medicación; esto evitará poner en riesgo la vida de los pacientes y su reingreso en estado crítico. El grado de precisión del algoritmo permitirá tomar decisiones basándonos en un modelo confiable.

### 2.2.7 Clasificación y Análisis Secuencial de Patrones para Mejorar la Eficiencia de Gestión y Proveer un Mejor Servicio Médico en Centros de Salud Públicos

El autor del presente estudio, cuyo título es “Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers”, fue Keunho Choi, PhD de la Universidad de Korea [CHO

2010]. El estudio tiene como objetivo determinar si un paciente volverá a asistir a consulta médica en una entidad de salud pública.

El conjunto de datos se recolectó del Centro de Salud Pública de Corea entre Enero del 2007 hasta Junio del 2008. Después de un proceso de eliminación de duplicados y valores nulos, se lograron obtener 7,057 muestras que entraron al modelo. Se escogieron 11 variables usando el evaluador de atributos “Gain Ratio”, el cual permite obtener los atributos reales más influyentes para el clasificador. La clase objetivo fue dividida en tres categorías: pacientes que volvían a consulta en 3 meses, en 6 meses y en 12 meses o más desde su primera consulta.

Para la etapa algorítmica se presentaron cinco algoritmos de clasificación: árboles de decisión, regresión lógica, redes neuronales, redes bayesianas y naive-bayesiano. Además se utilizó el método de validación cruzada partiendo el conjunto de datos en cinco partes. Esto significa que cada parte sirvió como conjunto de prueba y el resto como un conjunto de entrenamiento para cada iteración hasta obtener la distribución que otorgue mejores resultados.

Finalmente, se pudo obtener una alta precisión promedio para todos los algoritmos utilizados, siendo el mejor el de regresión lógica con 74.42%. Los otros algoritmos obtuvieron: árbol de decisión un 73.19%, redes neuronales un 74.05%, redes bayesianas un 74.3% y naive-bayesiano un 72.02%. Esto significa que un modelo predictivo puede ser eficientemente implementando ya que se obtienen valores de alta precisión en la clasificación. Los administradores de la entidad de salud podrán realizar acciones para motivar a los pacientes a continuar con sus consultas pendientes.

### 2.2.8 Factores que Afectan la Deserción Temprana y el Curso del Tratamiento Tardío de un Tratamiento Antidepresivo de Depresión en Circunstancias Naturales

Este estudio titulado “Factors affecting early attrition and later treatment course of antidepressant treatment of depression in naturalistic settings: An 18-month nationwide population-based study” fue realizado por Yi-Ju Pan, Shi-Kai Liu y Ling-Ling Yeh con el objetivo de identificar los factores influyentes en el abandono temprano del tratamiento para la depresión en el hospital Fast Eastern en Taiwán [PAN 2013].

Para este estudio participaron aproximadamente 216,557 pacientes mayores de edad que habían recibido medicación para la depresión. Se utilizaron, en total, 11 variables correspondientes a datos clínicos y socio-demográficos. Por último, los pacientes fueron divididos en tres categorías: los pacientes que no abandonan el tratamiento, los pacientes que regresan al tratamiento después de un tiempo y los pacientes que abandonan al tratamiento.

Con respecto al análisis, se utilizaron métodos observacionales y estadísticos. Entre los métodos estadísticos utilizados se encuentran: chi-cuadrado para las variables categóricas y ANOVA para las variables continuas. Primero, se optó por realizar un análisis multivariado para encontrar diferencias significantes entre las características y escoger las más importantes. Las características resultantes sirvieron para el modelo de regresión logística, el cual permitió identificar los factores influyentes en las tres categorías de la clase objetivo mencionadas anteriormente.

Los resultados mostraron que los casos de mayor abandono (27.9% del total de la población) correspondían a pacientes con las siguientes características: jóvenes, de sexo femenino, con depresión moderada, en consultas ambulatorias y los que presentaban alguna enfermedad física. Esto permitió obtener un patrón para cada grupo de pacientes y las relaciones con el abandono del tratamiento, y a partir de ello, comprender claramente el escenario del problema.

### 2.2.9 Otros Trabajos Relacionados

a) **Analysis of heart diseases dataset using neural network approach:** Este estudio tuvo como objetivo implementar un clasificador que utilice redes neuronales para predecir el grado en el cual se encuentra una enfermedad cardiaca. Los resultados mostraron un alto porcentaje de eficiencia: 90.6% y 94% para el modo de una capa y multicapas respectivamente; esto demuestra que el clasificador es bueno para resolver problemas en escenarios donde la información está muy distorsionada y no es lineal [USH 2011].

b) **Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques:** El objetivo de este estudio fue predecir los cargos hospitalarios para pacientes con cáncer utilizando redes neuronales y árboles de decisión.

Los análisis se realizaron separando a la población en dos grupos: total de pacientes y pacientes con seguro. Los resultados mostraron que el algoritmo de redes neuronales fue superior en todas las iteraciones, con un 98.7% de precisión como la más alta [KAN 2009].

c) **Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking:** El objetivo de este estudio fue determinar el mejor modelo de clasificación que permita predecir el abandono de un cliente en la Banca Minorista. Los algoritmos utilizados fueron: redes neuronales, naive bayesiano, clasificador K-NN (vecino más cercano), arboles de decisión, regresión logística y reglas de asociación. Los resultados mostraron que un árbol de decisión era el algoritmo ideal para este escenario debido a que tenía una alta precisión, 84.4%, y un tiempo de ejecución bajo (dos segundos), en comparación con los otros algoritmos [LIF 2011].

**Tabla 2.1**. Cuadro de Comparación del Estado del Arte. Elaboración propia.

| Título de los trabajos | Autores (País-Año) | Objetivo | Conjunto de Entrada | Algoritmos utilizados | Resultados o Eficiencia |
|---|---|---|---|---|---|
| 1) Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients | Peyman Rezaei Hachesu, Maryam Ahmadi, Somayyeh Alizadeh, Farahnaz Sadoughi (Iran-2013) | Predecir el tiempo de duración de los tratamientos para paciente cardiacos. | -2,065 muestras<br>-36 variables | 1) SVM<br>2) Algoritmos Combinados<br>3) Árbol de decisión<br>4) Redes neuronales | Eficiencia:<br>1) 96.4%<br>2) 95.9%<br>3) 83.5%<br>4) 53.9% |
| 2) Predictors of attrition from a pediatric weight management program | Meg Zeller, Shelley Kirk, Randal Claytor, Philip Khoury, Jennifer Grieme, Megan Santangelo, Stephen Daniels (EE.UU-2004) | Determinar los factores principales que llevan a abandonar el tratamiento para la obesidad en menores de edad. | -212 muestras | 1) ANOVA | Resultados:<br>-55% de pacientes abandonaron el tratamiento<br>-Los factores influyentes fueron: edad, autoestima y medicación. |
| 3) Using Data Mining to Describe Long Hospital Stays | Verónica Gómez, José E. Abásolo (Colombia-2009) | Predecir el tiempo de hospitalización y sus factores influyentes. | -68,732 muestras<br>-7 variables | 1)Reglas de asociación | Resultados:<br>-Los factores influyentes fueron: sexo y medicación. |
| 4) Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients | Youn-Jung Son, Hong-Gee Kim, Eung-Hee Kim, Sangsup Choi, Soo-Kyoung Lee (Corea-2010) | Determinar los factores críticos que influencian en el compromiso de un paciente por cumplir su medicación. | -76 muestras<br>-11 variables | 1) SVM | Eficiencia:<br>1) 77.63% |
| 5) Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers | Keunho Choi, Sukhoon Chung, Hyunsill Rhee, Yongmoo Suh (Corea-2010) | Predecir si un paciente volverá a asistir a su consulta médica. | -7,057 muestras<br>-11 variables | 1) Regresión logística<br>2) Redes bayesianas<br>3) Redes neuronales<br>4) Arbol de decisión<br>5) Naive bayesiano | Eficiencia:<br>1) 74.42%<br>2) 74.3%<br>3) 74.05%<br>4) 73.19%<br>5) 72.02% |

| | | | | | |
|---|---|---|---|---|---|
| 6) Factors affecting early attrition and later treatment course of antidepressant treatment of depression in naturalistic settings: An 18-month nationwide population-based study | Yi-Ju Pan, Shi-Kai Liu, Ling-Ling Yeh (Canada-2013) | Determinar los factores influyentes en el abandono temprano del tratamiento para la depresión. | -216,557 muestras<br>-11 variables | 1) ANOVA<br>2) Regresión logística | Resultados:<br>-27.9% de casos de abandono.<br>-Los factores influyentes fueron: edad, sexo, tipo de depresión, tipo de consulta y enfermedad actual. |
| 7) Analysis of heart diseases dataset using neural network approach | K. Usha Rani (India-2011) | Predecir el grado en el cual se encuentra una enfermedad cardiaca. | -414 muestras<br>-13 variables | Redes neuronales:<br>1) Multicapa<br>2)Una capa | Eficiencia:<br>1) 94%<br>2) 90.6% |
| 8) Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques | Jin Oh Kang, Suk-Hoon Chung, Yong-Moo Suh (Corea-2009) | Predecir los cargos hospitalarios para pacientes con cáncer. | -605 muestras<br>-65 variables | 1) Redes Neuronales<br>2)Árbol de decisión | Eficiencia:<br>1) 98.7%<br>2)Los factores influyentes fueron: duración de la estancia, número de operaciones, grupo de tratamiento |
| 9) Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking | Fan Li , Juan Lei , Ying Tian , Sakuna Punyapatthanakul, Yanbo J. Wang (China-2011) | Predecir el abandono de un cliente en Banca Minorista. | -4,000 muestras<br>-12 variables | 1) Árbol de decisión<br>2) Reglas de asociación<br>3) Redes neuronales<br>4)Clasificador K-NN<br>5) Reglas de asociación<br>6) Naive bayesiano | Eficiencia:<br>1) 84.4%<br>2) 84.7%<br>3) 83.95%<br>4) 81.5%<br>5) 79.3%<br>6) 61.8% |

### 2.2.10 Conclusiones sobre el estado del arte

De acuerdo a los resultados de los trabajos mencionados anteriormente, se puede concluir que es factible la implementación de un modelo predictivo de buena precisión y la obtención de sus factores influyentes, lo cual servirá de apoyo en la toma de decisiones de un hospital. Además, se ha podido identificar una gran cantidad de algoritmos con altos resultados de rendimiento, como el SVM, Árbol de decisión, Redes neuronales y Reglas de Asociación, que permitirán hacer el proceso de Minería de Datos iterativo y poder conseguir el mejor resultado para el modelo.

A diferencia de las investigaciones presentadas, el proyecto académico de fin de carrera estará también enfocado en la integración de la información desde las diferentes fuentes de origen. La etapa de pre procesamiento tendrá diferencias debido a que utilizarán distintos métodos para la normalización de los datos, manejo de missing values y detección de valores anómalos. Se planea utilizar varios algoritmos en un proceso iterativo para obtener el que provea mejor resultado como se ha observado en la investigación realizada por Maryam Ahmadi. Sin embargo, la diferencia radica en el enfoque del problema: el proyecto de fin de carrera se centra mayormente en el tratamiento (proceso largo para curar una enfermedad) y el descubrimiento de los factores de abandono, mientras que en la mayoría de los estudios presentados, en el tiempo de hospitalización (ocupación de los cuartos del hospital). Las investigaciones realizadas por Meg Zeller (2.5) y Shi-Kai Liu (2.9), buscan resolver un problema similar al del proyecto de fin de carrera; la diferencia, en este caso, será la técnica de Minería de Datos que se utilizará contra los métodos estadísticos usados en dichas investigaciones.

Finalmente, se puede observar que las investigaciones presentadas en el Estado del Arte permiten apoyar los diferentes objetivos específicos que se desean resolver en el proyecto académico. Por lo tanto, servirán de guía para el desarrollo de la solución planteada.

# Capítulo 3: Creación del DataMart de Pacientes

En este capítulo se tratarán los resultados obtenidos para el primer objetivo específico. El primer objetivo específico consiste en diseñar un modelo de base de datos que permita integrar las diferentes fuentes de información de la entidad de salud de una forma automatizada. Para este objetivo se realizará la creación de un Datamart con los datos necesarios para el modelo y se implementará un proceso ETL que permita la integración automatizada de la información.

## 3.1 Resultado Esperado 1: Creación del modelo de DataMart con los datos necesarios

De acuerdo con la problemática, la información generada en las instituciones de salud es de un nivel complejo debido a su gran volumen y a las diferentes fuentes de donde puede provenir. Esto podría dificultar la labor de análisis de los investigadores o del área de estadística, los cuales podrían necesitar la extracción de información para los diversos estudios de forma precisa y oportuna.

Como parte de la etapa de pre-procesamiento en un proceso de Minería de Datos, es importante que la información pueda estar organizada e integrada de tal forma que sea fácil obtener las variables que servirán para entrenar el modelo [MEH 2011]. Además, esto permitirá que sea más sencillo el proceso periódico a través de los ETL's que se implementarán en el siguiente resultado.

Actualmente, el Instituto Nacional de Enfermedades Neoplásicas no cuenta con un Data WareHouse; por lo tanto, se hace necesaria la creación de un Datamart que pueda integrar la información de las diferentes tablas de la Base de Datos de origen. Por tal motivo, se han realizado reuniones con los encargados del área de sistemas y doctores involucrados para conocer detalladamente los procesos que se realizan en el hospital, delimitar los que interactúan con variables que podrían tener influencia en el modelo, identificar las tablas fuentes donde se extraerá la información y, finalmente, definir el modelo más adecuado para trabajar dichas variables.

En el anexo 10.1, se puede observar el diagrama del proceso general que se realizará para la extracción de la información y creación del modelo de datos. En este punto se

debe aclarar que el DataMart servirá como fuente de datos para el proyecto académico de fin de carrera y para un proyecto con el INEN. Lo ideal es que el DataMart pueda integrar ambas soluciones ya que tienen elementos en común.

A continuación, se presenta el modelo de base de datos del DataMart y se explicará su desarrollo y las principales entidades involucradas, así como la importancia que tendrán para implementar el modelo predictivo. Adicionalmente, se puede encontrar el diccionario de datos con tamaño, tipo de dato y la descripción de cada atributo en el anexo 5.2.

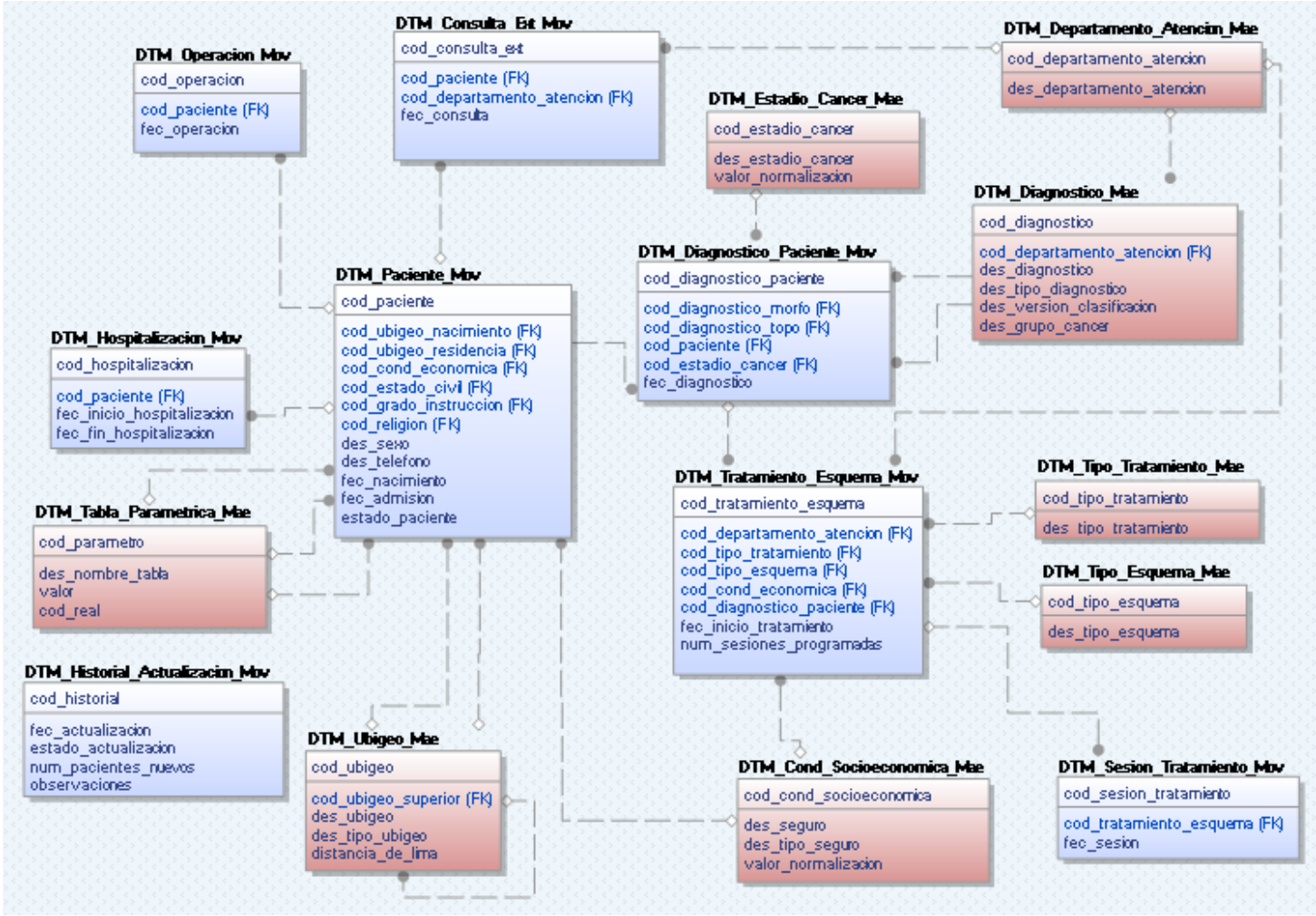


**Figura 3.1.** Modelo del DataMart. Imagen propia.

### 3.1.1 Modelo del DataMart

Como se mencionó en la sección de herramientas, Erwin será el software que se utilizará para el modelado de las tablas del DataMart.

Un DataMart deberá estar enfocado en el alcance del problema que se pretende resolver. Por tal motivo, el siguiente paso es identificar las variables más importantes, y a partir de ello, escoger el esquema más apropiado para el modelo del DataMart que pueda apoyar la problemática planteada.

### 3.1.2 Descripción de las entidades influyentes en el modelo

Las entidades más importantes para poder entender los procesos son las siguientes:

**a) Paciente**

Esta entidad contiene los datos principales del paciente. Al estar trabajando en una institución bajo regulaciones de protección de la información, se debe mantener la información personal del paciente como privada, por lo cual se ha decidido ocultar datos como nombres, apellidos, DNI, entre otros. La forma en que se reconocerá al paciente será a parir de su historia clínica, dato que solo es de conocimiento dentro de la institución. Durante el proceso de análisis con los miembros de la institución, se ha podido apreciar que las variables como religión, lugar de nacimiento, lugar de residencia y estado civil pueden jugar un papel muy importante en el proceso de clasificación, incluyendo a las variables comunes como sexo, diagnostico, condición socioeconómica (tipo de seguro), nivel del cáncer (estadío), entre otros.

**b) Tratamiento_Esquema**

Esta entidad contiene la información de cada tratamiento que debe recibir el paciente. De acuerdo al análisis, se pudo identificar que dependiendo del tipo de cáncer y el estado del paciente, se otorga un esquema de tratamiento, el cual varía en intensidad y podría ser una fuerte causa de abandono para pacientes que no puedan soportar fuertes tratamientos. Por este motivo, también se almacenan los datos del diagnóstico del paciente para el tratamiento en mención. Además, se puede obtener la condición socioeconómica del paciente al momento de recibir el tratamiento, debido a que esta

puede cambiar y alterar los resultados del modelo predictivo. Finalmente, se pueden encontrar datos importantes como la duración del tratamiento, número de sesiones programadas y recibidas, las cuales podrían afectar si es que el tratamiento es muy prolongado; y el cumplimiento, el cual representa la variable objetivo.

**c) Sesion_Tratamiento**

Esta entidad contiene las sesiones por cada tratamiento que ha recibido el paciente. Esto permitirá identificar finalmente que pacientes han asistido a la mayor parte de sus sesiones programadas.

**d) Diagnostico_Paciente**

Esta entidad contiene los diagnósticos por paciente. Cabe mencionar que para el cáncer se realizan dos tipos de diagnósticos: Morfológico (enfocado en la forma del cáncer) y Topográfico (enfocado en el lugar donde reside el cáncer). Se utilizará la clasificación basada en el CIE-10. Ambos datos son de vital importancia porque, de acuerdo a las estadísticas del hospital, existen tipos de cáncer que requieren de complicados procedimientos para poder ser tratados, lo cual termina conduciendo a que se pueda producir el abandono. Otro factor importante es el estadío del cáncer, tanto al inicio como durante el tratamiento, ya que puede determinar si es que el tratamiento es efectivo.

**e) Otras Entidades**

Entre otras entidades que podrían proporcionar variables que apoyen a una mejor predicción del modelo se encuentran:

- Operación: El número de operaciones que ha tenido el paciente en los últimos meses.
- Hospitalización: El número de hospitalizaciones en los últimos meses podría dar una imagen de mayor gravedad de la enfermedad, así como la duración.
- Consulta Externa: El número de consultas que ha tenido en los últimos meses.

### 3.1.3 Tipo de DataMart

El tipo de DataMart por el cual se ha optado es un esquema de copo de nieve, el cual resulta de la expansión de las dimensiones del modelo estrella. Esto permite que el

modelo pueda ser más normalizado y entendible para los miembros de la organización debido a la baja cardinalidad de cada campo. Como ejemplo se pueden observar las tablas de “Estadio Cáncer” y de “Diagnostico Cáncer”, las cuales presentan atributos que, en un modelo estrella, se encontrarían en la tabla del “Diagnostico del Paciente”, lo cual la haría poco entendible y sobrecargada. Además, como otro beneficio de este modelo se puede mencionar una reducción en el espacio total que ocuparía el modelo, lo cual es un aspecto muy importante para el uso de los recursos informáticos dentro de la institución. Por estas razones, el esquema copo de nieve es el más adecuado para el modelo desarrollado [IBM 2005].

### 3.1.4 Estándares del modelo

Los estándares del modelo se encuentran basados en algunos estándares existentes en la Base de Datos del INEN:

- Las tablas tendrán el prefijo “DTM_” para poder identificar las tablas que pertenecen al modelo de DataMart y diferenciarlas de las existentes en la Base de Datos.
- Las tablas maestras tendrán el sufijo “_Mae” .
- Las tablas transaccionales tendrán el sufijo “_Mov”.
- Las llaves primarias tendrán el prefijo “cod”
- Los campos que representen nombres o denominaciones tendrán el prefijo “des”
- Los campos que representen fechas tendrán el prefijo “fec”.
- Los campos que representen cantidades tendrán el prefijo “num”.
- Los campos tendrán como máximo dos palabras en minúscula, lo cual permitirá identificarlos de una forma más exacta.
- Los campos utilizarán sub-guiones para separar identificadores y palabras.

### 3.1.5 Conclusiones

De acuerdo a lo planteado anteriormente, la creación de un DataMart, el cual pueda integrar y organizar la información para su posterior análisis, es muy importante y

forma parte de la etapa de pre-procesamiento de Minería de Datos para poder refinar los datos [MEH 2011].

Se ha optado por el modelo que sea adecuado, tanto para las condiciones actuales de la base de datos del INEN y para el problema que se pretende resolver en el proyecto. Con respecto al abandono del tratamiento, se ha planteado tomar solo los casos de quimioterapia como se ha mencionado en el alcance; sin embargo, el modelo permitirá ser adaptado para todos los tipos de tratamiento si es que así se desea en el futuro.

Cabe mencionar que el propósito del DataMart es, no solo servir de apoyo a la toma de decisiones con respecto al abandono de tratamiento, sino que también pueda ser de apoyo para obtener conocimiento relacionado con otros problemas en el futuro. El DataMart también servirá como una herramienta de estadística y de reportes dinámicos para los investigadores y doctores, lo cual no formará parte del proyecto de Tesis. Por tal razón, se ha optado por que el modelo de base de datos pueda integrar la información necesaria para ambos proyectos e incluso servir de apoyo para otros problemas que se desee analizar.

## 3.2 Resultado Esperado 2: Proceso ETL que permita la integración automatizada de la información

Debido a la excesiva carga de funciones que existen en una entidad de salud pública, es de vital importancia que los procesos que se planteen sean lo menos manuales posibles. Más aún si es que para la carga de información del DataMart, se tendrían que realizar varias actividades que impliquen extraer información de las diversas fuentes involucradas en el modelo.

Por esta razón, el proceso de ETL permitirá realizar dichas actividades de forma automatizada utilizando la herramienta “Data Integration” del conjunto de soluciones para análisis de negocios conocida como Pentaho. Esta herramienta posee una interfaz fácil de entender y una forma sencilla para estructurar los procesos de extracción, transformación y carga. Otra función importante es que se puede dar seguimiento al proceso completo para determinar que cada actividad se realice de acuerdo a lo que se había planeado [PEN 2014].

Como se puede apreciar en la figura 10, modelo del DataMart, se han mostrado de diferentes colores las tablas transaccionales y las maestras. Es importante separarlas debido a que en el proceso ETL, las tablas maestras comúnmente cargarán información la primera vez que se ha ejecutado el proceso, mientras que las tablas transaccionales serán actualizadas periódicamente, de acuerdo al tiempo que sea acordado. En el anexo 10.3 se puede encontrar la descripción de los principales procesos que se realizarán para la carga del DataMart.

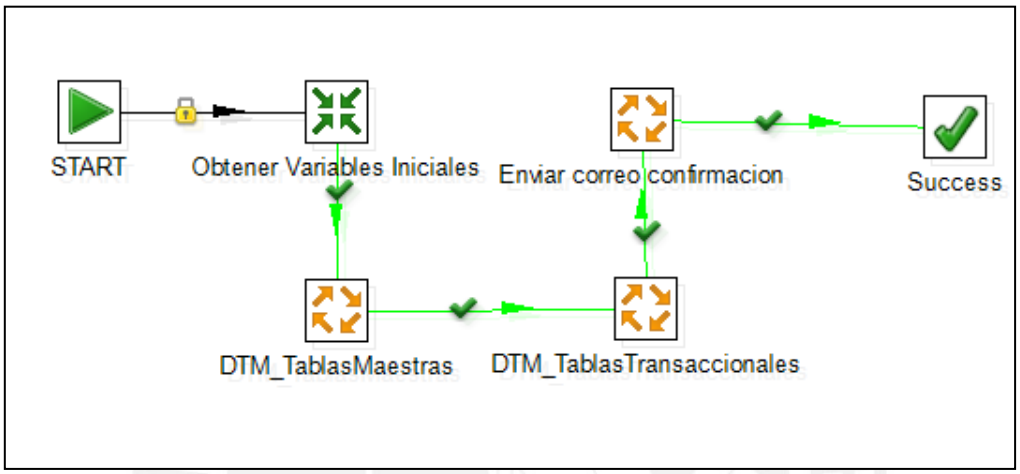


**Figura 3.2.** Proceso ETL general. Imagen propia

### 3.2.1 Carga de tablas maestras

Con respecto a los procedimientos para realizar la carga de datos de las tablas maestras, se evaluará si es que no existen actualizaciones anteriores, de lo contrario el procedimiento permitirá actualizar automáticamente las tablas con las instancias nuevas. De acuerdo con el análisis realizado con los miembros de la institución, se ha decidido solo tener en cuenta el ingreso de nuevas instancias en las tablas maestras, más no la modificación o edición de alguna de ellas. A continuación se muestra la transformación para cargar los datos a la tabla de diagnósticos como ejemplo, donde se tiene en cuenta la actualización de las tablas maestras:

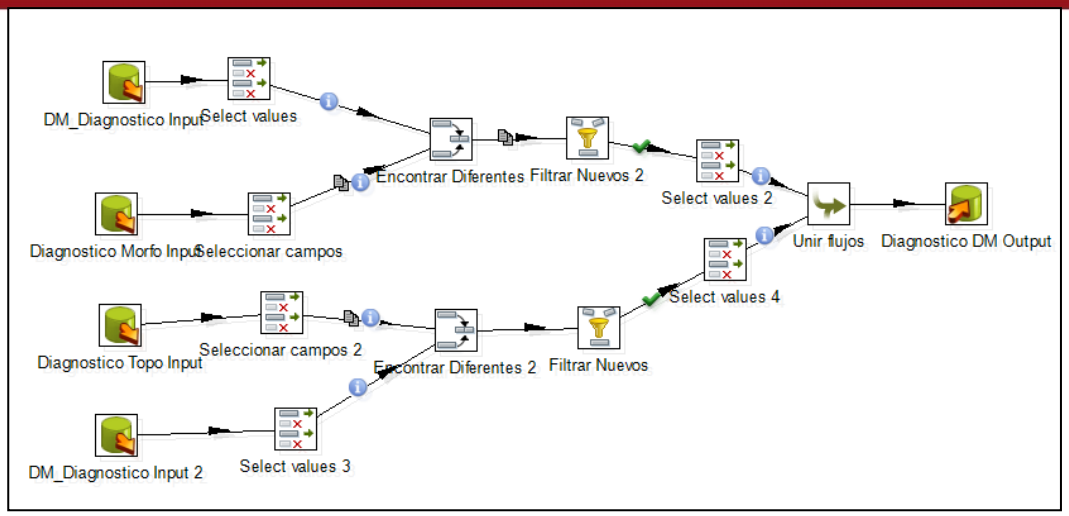


**Figura 3.3.** Transformación de la tabla maestra "DMT_Diagnostico_MAE". Imagen propia

### 3.2.2 Carga de Tablas Transaccionales

Con respecto a las tablas transaccionales, estas se actualizarán periódicamente, de acuerdo al tiempo que se llegue a consensuar con los miembros de la institución. Antes de iniciar la carga de datos se obtiene como parámetro la última fecha de actualización del DataMart y la fecha actual, con el fin de que la carga de nuevos datos a las tablas transaccionales se pueda realizar, a partir del último día que fue cargado hasta el día anterior a la fecha de actualización. Se realizó la creación de transformaciones por cada tabla que sea cargada para que se pueda mantener organizado y entendible el proceso, ya que se puede requerir mantenimiento en el futuro.

Finalmente, es importante que los miembros de la institución puedan estar al tanto del éxito de la actualización del DataMart para poder realizar sus actividades de investigación y estadísticas. Por tal motivo, como un proceso final del flujo de carga del DataMart, se realizará la confirmación del proceso enviando un correo a los miembros de la institución a través del correo del INEN.

### 3.2.3 Conclusiones

Como se mencionó, el proceso automatizado de carga del DataMart es muy importante debido a que permitirá liberar de trabajo adicional a los miembros de las

áreas de Sistemas y Estadística, y así mismo, poder dar un seguimiento de cada proceso que sea ejecutado.

Otro factor importante es que la estructura de los procesos pueda ser entendible para los miembros de la institución, en caso se desee actualizar o adaptar el modelo en el futuro. Pentaho ofrece herramientas bastante intuitivas y fáciles de aprender, las cuales facilitan la creación y mantenimiento del proceso ETL. Como se ha mostrado en las figuras de los procesos de carga, tanto de las tablas maestras como las transaccionales, se ha buscado mantener una estructura organizada y descripciones adecuadas de cada actividad para que este fin pueda ser cumplido.

Para concluir, es importante tener en cuenta que cada actualización del proceso debe ser comunicada a los involucrados para que puedan realizar sus labores, por lo cual se realizará el envío de correos para la confirmación del proceso.

# Capítulo 4: Pre-Procesamiento del modelo de Minería de Datos

En este capítulo se explicarán los resultados obtenidos para el segundo objetivo específico, el cual consiste en identificar los escenarios donde las variaciones y anomalías puedan afectar de manera negativa en el modelo, y realizar las acciones correctivas para obtener los resultados de precisión que se tienen como meta. Para este objetivo se realizará el método de tratamiento de valores nulos, el método de normalización de datos y, finalmente, el algoritmo de detección y eliminación de outliers o valores anómalos. Estos procedimientos permitirán refinar la información como parte de la etapa de pre procesamiento.

## 4.1 Resultado Esperado 3: Método de tratamiento de valores nulos o inconsistentes

En el caso de las entidades de salud públicas que manejan grandes cantidades de información es muy común observar que existen campos con valores nulos. Esto puede ser generado por omisión del propio sistema, pérdida de datos, incoherencias e incluso omisión por pedido del propio paciente [MEH 2011]. Por esta razón es muy importante determinar qué mecanismos se utilizarán para aplicar un valor apropiado a los campos nulos basándose en las relaciones de cada muestra y la experiencia que puedan brindar los miembros de la entidad de salud.

En la tabla 4.1, se puede observar los resultados de la cantidad de valores nulos para cada variable, así como el porcentaje que representan. A continuación se procederá a explicar los mecanismos que se utilizaron para poder lidiar con estos valores.

**Tabla 4.1**. Cuadro de resultados de valores nulos. Elaboración propia.

| Variables | Total Nulos | Total Muestras | Porcentaje Aprobación |
|---|---|---|---|
| Condición Económica | 0 | 4425 | 100.0% |
| Estado civil | 0 | 4425 | 100.0% |
| Lugar nacimiento | 0 | 4425 | 100.0% |
| Lugar residencia | 0 | 4425 | 100.0% |
| Religión | 4291 | 4425 | 3.03% |
| Nivel Instrucción | 2534 | 4425 | 42.73% |
| Sexo | 0 | 4425 | 100.0% |
| Edad | 0 | 4425 | 100.0% |
| Primera Consulta | 0 | 4425 | 100.0% |
| Diagnostico topográfico | 63 | 4425 | 98.58% |
| Estadio_Cáncer | 796 | 4425 | 82.01% |
| Numero operaciones | 0 | 4425 | 100.0% |
| Numero hospitalizaciones | 0 | 4425 | 100.0% |
| Brecha | 0 | 4425 | 100.0% |
| TratamientoAbandona | 0 | 4425 | 100.0% |

### 4.1.1 Eliminación de muestras o variables con exceso de valores nulos

Para poder tener un conjunto de datos limpio, es necesario que la mayor parte de las muestras sea consistente y no existan demasiados valores nulos, de lo contrario no se podrían formar patrones definidos que permitan una predicción acertada del modelo.

Como primer paso, se debe analizar que cada muestra tenga la mayor cantidad de campos con valores completos, ya que por el mismo motivo, estas muestras podrían distorsionar el modelo [HOR 2007]. En este caso se ha escogido el valor de 80% para tomar en cuenta o no dichas muestras. Este proceso deberá realizarse continuamente para que el conjunto de datos final pueda tener datos coherentes.

Como segundo paso, es importante determinar el porcentaje de valores nulos que se tienen por variable. Esto es importante ya que se podrá determinar si una variable contribuirá con el modelo o no, al definir su comportamiento para cada grupo [HOR 2007]. En el caso de que existan muchos valores nulos para esa variable, no se podrá

diferenciar un comportamiento distinto para cada grupo ya que todo se concentrará en los valores nulos. Como un límite, se ha planteado el valor de 80% para decidir si es que la variable será tomada en cuenta en el modelo. Las variables, religión y nivel de instrucción, que fueron eliminadas del modelo se pueden observar en la tabla 4.

Se ha identificado en documentación revisada que los porcentajes de aprobación mencionados varían dependiendo de la problemática, por lo cual se ha optado por determinar estos valores junto a los miembros de la institución [HAC 2013] y [CHO 2010].

### 4.1.2 Reemplazar los valores nulos con valores probables o esperados

Una vez que las variables o muestras con exceso de valores nulos han sido eliminadas, es necesario determinar cuáles son los mecanismos que se utilizarán para los campos que aún permanecen como nulos. Debido a que el conjunto de datos de entrada para el modelo necesita, en su mayoría, datos numéricos, estos campos no pueden ser nulos, por lo cual se les debe asignar algún valor dependiendo de la experiencia de los miembros de la entidad de salud [MEH 2011]:

- En el caso del estadio, se asignará el valor de 2 a todos los campos nulos. De acuerdo a los doctores, estos pacientes, en su mayoría, aún no han sido diagnosticados y se les debería considerar de una gravedad intermedia.
- En el caso del diagnóstico topográfico, al ser un valor nominal, se marcará como “No diagnosticados” los campos nulos.

### 4.1.3 Conclusiones

Se han aplicado los mecanismos para el tratamiento de valores nulos que permitan eliminar muestras y variables con un exceso de valores nulos, los cuales podrían distorsionar el modelo, y luego reemplazar los valores que aún permanecen incompletos con valores apropiados. Por esta razón se han eliminado dos variables con exceso de valores nulos en el modelo (nivel de instrucción y religión) y se ha decidido como reemplazar a las demás variables con valores nulos. Esta etapa es de vital importancia ya que se necesita tener un conjunto de datos limpio y con datos consistentes para obtener una mayor precisión en los resultados del modelo.

## 4.2 Resultado Esperado 4: Método de normalización de datos

Muchos algoritmos de Minería de Datos necesitan conjuntos de datos con un tipo de transformación especial, con el fin de que estos puedan ser consistentes con el algoritmo escogido, y a su vez, obtener resultados con mayor precisión [MEH 2011]. Por ejemplo, algoritmos de clasificación como el kNN (k vecinos más cercanos), deberían tener los datos como valores numéricos debido a que utilizan cálculos basados en distancia, en caso contrario, las variables de cadena de caracteres tomarán valores que pueden no ser los más precisos. Además, algunos algoritmos aumentan su precisión y rapidez cuando la cantidad de valores distintos por variable es reducida, por lo cual un método de suavización de datos es adecuado para optimizar el proceso general.

En el anexo 10.4 se puede observar que procedimientos de normalización ha seguido cada variable del modelo.

### 4.2.1 Transformación de cadena de caracteres

Debido a que algunos algoritmos serán basados en distancia, es necesario que algunas de las variables de tipo cadena de caracteres puedan transformarse en tipo numéricas. Por tal motivo, es importante distinguir las variables cualitativas nominales y ordinales para aplicar el procedimiento respectivo [MEH 2011]:

- Variables Nominales: Debido a que en estas variables no existe jerarquía, una conversión de tipo de datos puede no tener un gran impacto en el modelo, por lo cual, no es necesario que sean cambiadas.
- Variables Ordinales: En este tipo de variables existe una relación de orden por lo cual se puede reemplazar la variable numéricamente, asegurando una mayor exactitud en el modelo.

**Tabla 4.2**. Cuadro de variables de cadena de caracteres.
Elaboración propia.

| Variables Nominales | Variables Ordinales (criterio de cambio) |
|---|---|
| Estado civil<br>Sexo<br>Diagnóstico topográfico | Condición Económica (tipo de seguro)<br>Estadío (gravedad del cáncer) |

### 4.2.2 Normalización Mínimo-Máximo

Este método de normalización permite que los valores de cada variable puedan estar dentro de un rango establecido que sea el mismo para todas, en este caso, el rango es [0,100]. Las ventajas de esta normalización es que permitirá que todas las variables tengan la misma capacidad de influir en el modelo [MEH 2011]. Por ejemplo, suponiendo que una variable tiene su rango de valores [30,150], será mucho más influyente al momento de la predicción que una variable con rango [15,20] en las técnicas basadas en distancia.

Por tal motivo, se ha escogido realizar el método de normalización Mínimo-Máximo para todos los campos con valores numéricos. Esto también incluye a variables ordinales que serán reemplazados por valores numéricos, con el fin de que el conjunto de datos de entrada pueda mostrar una mayor consistencia en el modelo.

A continuación se muestra la fórmula que se utilizó para el procedimiento de conversión:

$$v'(i) = \frac{(v(i) - \min(v)) * 100}{(\max(v) - \min(v))}$$

donde:

i: número de variable.

v(i): valor originar del campo de la variable i.

v'(i): valor normalizado del campo de la variable i que entrará en el modelo.

En la figura 4.1, que se mostrará a continuación, se puede observar un ejemplo del método de normalización min-max y como quedaría, finalmente, la variable después de la normalización.

| | Lugar Residencia-Sin Normalización | Lugar Residencia-Normalizado |
|---|---|---|
| **Min** | 0 | 0 |
| **Max** | 1380 | 100 |
| **1** | 310 | 22 |
| **2** | 0 | 0 |
| **3** | 1050 | 76 |
| **4** | 130 | 9 |

$$v(1) = 22 = \frac{(310 - 0) * 100}{1380 - 0}$$

**Figura 4.1.** Ejemplo de normalización Min-Max. Imagen propia.

### 4.2.3 Suavización

Como actividad final de la etapa de normalización se debe procurar que no exista una gran cantidad de datos distintos para cada variable. Esto podría generar que los algoritmos tomen mayor tiempo y sean menos precisos, ya que diferencias menores no harán un gran impacto en las futuras decisiones del modelo [MEH 2011]. Por tal motivo se han realizado los siguientes métodos de suavización:

Con respecto a las variables de cadena de caracteres, se realizarán grupos, de acuerdo a la experiencia de los doctores, para clasificar los distintos tipos de diagnósticos y lugares de nacimiento y residencia:

- La variable diagnóstico topográfico tiene 1190 valores distintos provistos por la CIE-10; cada uno de estos será agrupado por el departamento de atención en el cual serán atendidos.
- En el caso de las variables lugar de nacimiento y lugar de residencia, serán agrupadas por departamentos y se tomará en cuenta la distancia desde dicho lugar hasta la ciudad de Lima. Dichas distancias se encontrarán en una hoja de

cálculo (para tener un indicador del esfuerzo que deben realizar los pacientes para llegar a la sede del INEN)

Las variables numéricas, después de haber sido normalizadas, tendrán varios decimales, lo cual las convertirá en varios valores diferentes. Por esta razón, un método simple de normalización es redondear el número a su valor entero.

### 4.2.4 Conclusiones

Se han realizado los métodos de normalización con el fin de que los datos crudos puedan refinarse y, ser consistentes con lo que cada algoritmo requiere. Esto implica una conversión de datos de las variables ordinales a números, generalmente, para los algoritmos basados en distancia, la normalización dentro de un rango idéntico para todas las variables numéricas y, finalmente, la suavización de los datos para reducir la cantidad de variables distintas.

Cabe resaltar que estos métodos varían dependiendo de la necesidad de cada algoritmo para la primera parte del proyecto, ya que aún no se cuenta con el algoritmo de mayor precisión. Una vez que el algoritmo más eficiente haya sido escogido, el proceso se realizará automáticamente y se utilizarán solo los métodos que sean necesarios.

## 4.3 Resultado Esperado 5: Algoritmo de detección y eliminación de valores anómalos

Cuando se realiza el análisis sobre grandes cantidades de información es muy frecuente encontrar muestras que no se encuentren dentro de los patrones del grupo al cual pertenecen. Más aún, cuando las muestras son datos que corresponden a personas ubicadas geográficamente a nivel nacional (lo cual implica una gran cantidad comportamientos distintos), se puede identificar una mayor variación entre los comportamientos de cada muestra dentro de un mismo grupo, lo que finalmente puede distorsionar la capacidad que presente el modelo cuando se desee predecir la continuidad en el abandono de tratamiento [MEH 2011].

Por tal motivo se ha escogido realizar un algoritmo de detección de valores anómalos que permita eliminar dichos valores y así poder delimitar los patrones dentro de cada grupo de característica que se pretende clasificar. Como resultado, el algoritmo generará una salida con las muestras que integrarán el conjunto de datos final del modelo. Este procedimiento solo se realizará para el conjunto que generará el modelo, no para el conjunto de pacientes a los cuales se buscará predecir su comportamiento.

Existen distintas técnicas enfocadas en la detección de valores anómalos, tales como los métodos estadísticos, los métodos basados en distancia y los métodos basados en modelo. El proyecto utilizará las técnicas basadas en modelo, las cuales forman parte del grupo que más se acerca a la forma en que el razonamiento del hombre podría distinguir las muestras inusuales. La característica de esta técnica es que define un conjunto de patrones e identifica los valores que se desvían de esos valores [MEH 2011].

### 4.3.1 Algoritmo kNN

El algoritmo que se ha implementado, dentro de las técnicas basadas en modelo, es conocido como el kNN o los k vecinos más cercanos, el cual pertenece al grupo de algoritmos “lazy learning”. Cabe destacar que este algoritmo es utilizado para muchos otros tipos de problemas, como clasificación, clusterización, regresión, entre otros, donde puede variar su tiempo de complejidad [MEH 2011].

Si bien el tiempo de un algoritmo kNN para detección de valores anómalos puede ser bastante alto, se ha escogido debido a que permite una fácil implementación y determinar los valores anómalos de forma parecida al razonamiento humano, lo cual es esencial para este tipo de escenarios que dependen del comportamiento, a diferencia de los algoritmos basados en distancias. Como se mencionará en el resultado esperado 8, la búsqueda de una mayor precisión en las predicciones del abandono de tratamiento es de vital importancia para comprobar que nuestro modelo será útil y efectivo para la toma de decisiones de la institución.

A continuación se puede apreciar el pseudocódigo general del algoritmo kNN para la detección de valores anómalos y el sub procedimiento para la elección de los vecinos más cercanos de cada muestra y, posteriormente, se explicarán algunos detalles importantes acerca de la elección de los parámetros utilizados

### 4.3.2 Pseudocódigo del algoritmo kNN

```
OUTLIERKNN (k: Entero, umbral: Entero, data: Modelo)

ArregloPuntaje:={vacío}

VecinosCercanos:=obtenerKVecinosPorMuestra(k,data)

//Calcular Puntaje
Para i := 1 .. data.numMuestras Hacer:
        ArregloPuntaje[i]:=asignarPuntaje(VecinosCercanos,i)

//Eliminar muestra del modelo
Para i := 1 .. data.numMuestras Hacer:
        Si Puntaje[i]<umbral Entonces:
                Eliminar data[i]

END OUTLIERKNN
```

**Figura 4.2.** Pseudocódigo del algoritmo kNN. Imagen propia.

Los procedimientos involucrados en el algoritmo se explicarán a continuación:

- obtenerKVecinosPorMuestra: Este procedimiento permitirá obtener los k vecinos más cercanos para cada muestra dentro de cada grupo de la función objetivo del modelo de datos, lo cual implica que solo se compararán las distancias euclidianas entre muestras que pertenezcan a la misma clase (continua o abandona).

- asignarPuntaje: Este procedimiento permitirá asignar un puntaje para cada muestra. Este puntaje representa la cantidad de veces que dicha muestra es vecina de otras de su misma clase, lo cual define si es que la muestra se encuentra lejos de la concentración por grupo. Si el valor del puntaje es bajo, se puede considerar como un valor potencialmente anómalo.

### 4.3.3 Parámetros

El algoritmo kNN contempla dos parámetros principales los cuales se han determinado basándose en las experiencias que se podrán observar en la sección 5.4.4:

- **k:** El valor de la variable k representa el número máximo de vecinos cercanos que se va a registrar por muestra. Para esto se escogerán los que tengan una distancia menor a la muestra que se está analizando. El valor que se ha establecido para k es 3, ya que es el valor recomendado en la mayoría de documentación revisada y ha sido el que mayor mejora de precisión ha dado en la evaluación de resultados.

- **umbral:** El valor de la variable umbral representa el límite mínimo de ocurrencias como vecino que debería tener una muestra para que se pueda considerar dentro del modelo. En caso la muestra tenga un puntaje menor al establecido como umbral, dicha muestra no será tomada en cuenta. Para el caso del valor de umbral, se ha determinado el valor de 2.

### 4.3.4 Conclusiones

Se ha utilizado el algoritmo kNN para la detección de valores anómalos debido a que su implementación no es muy complicada y permite obtener una mayor exactitud de los valores anómalos, ya que es un algoritmo que se acerca a la forma en que los humanos podrían detectar estos elementos, pero enfocado para una gran cantidad de dimensiones. Con respecto a los parámetros utilizados, estos han sido escogidos basándose en documentación revisada y con el fin de que el conjunto de datos final pueda mantener su integridad y volumen para cada grupo de la función objetivo.

Finalmente, este procedimiento forma parte de la última etapa de pre-procesamiento del conjunto de datos; es decir, la salida generada servirá como entrada para cada uno de los algoritmos escogidos dentro del proceso de Minería de Datos, por lo que se generará el archivo de entrada adecuado (.arff) para el siguiente paso.

# Capítulo 5: Evaluación de los algoritmos

En este capítulo se trataran los resultados obtenidos para el tercer objetivo específico, el cual consiste en evaluar los algoritmos de clasificación a través de un número de iteraciones del conjunto de datos, con el fin de determinar cuál de los algoritmos tiene la mayor precisión y tomarlo en cuenta para el proceso automatizado de Minería de Datos. Para este objetivo se realizará la definición del modelo de datos de entrada para cada algoritmo, el análisis de los algoritmos escogidos para el proceso iterativo, la evaluación de los algoritmos a través del método de validación cruzada y, finalmente, el informe con los resultados de precisión para comprobar la mejora de cada método y de cada uno de los algoritmos. Esto permitirá identificar cual es el algoritmo que funciona mejor para las condiciones del conjunto de datos.

## 5.1 Resultado Esperado 6: Modelo de datos para cada algoritmo.

Una vez que la etapa de pre procesamiento ha concluido, es necesario definir los conjuntos de datos que entrarán para cada algoritmo. Como se ha venido mencionando en secciones anteriores, cada algoritmo tendrá una mayor precisión si el conjunto de datos que servirá como entrada tiene las características apropiadas y es consistente. En el caso de algoritmos como bayesiano o SVM, los datos de entrada deben ser numéricos y normalizados con el fin de que puedan ser los más adecuados; por otro lado, para algoritmos de árboles, datos de cadena de caracteres, con pocos valores distintos, permiten una mejor clasificación [MEH 2011].

### 5.1.1 Tipo de Archivo

Como se mencionó en la sección de herramientas, se ha escogido Weka debido a que tiene las funciones necesarias para poder facilitar el proceso de Minería de Datos y permite usarse como librería para programación en el lenguaje Java.

Weka recibe los datos de entrada en un archivo de extensión (.arff). Este tipo de archivo contiene las instancias o muestras de un modelo de datos que comparten un conjunto de atributos. Este archivo tiene la siguiente estructura [WEK 2014]:

- **Cabecera:** Contiene el nombre de la relación, los atributos con su respectivo tipo de dato y el atributo objetivo con los posibles valores de clasificación. Los atributos pueden ser de dos tipos
    - Numéricos
    - Cadena de caracteres: En la estructura del archivo aparecen los valores distintos que se encuentran por cada variable de este tipo
- **Datos:** Cada línea representa una instancia con los valores de sus atributos separados por comas

Finalmente, debido a que el proceso se va a realizar automáticamente, se requiere que el archivo (.arff) pueda ser creado una vez que el pre-procesamiento haya concluido. Por tal motivo, se utilizó un plugin de Pentaho, cuyo nombre es “Arff Output”, que convertirá la salida de la etapa de detección y eliminación de valores anómalos, descrita en el capítulo 4, en un archivo de tipo (.arff) que permita explotar las funciones de Weka [PEN 2014].

En la figura 5.1 se puede observar un fragmento del modelo de datos de tipo arff que se utilizará para la evaluación del modelo con los algoritmos de clasificación y luego, en la tabla 5.1, se puede observar la definición de cada variable considerada.

```
@relation TratamientoINEN
@attribute 'Condicion Economica' numeric
@attribute 'Estado Civil' {CASADO/A,CONVIVIENTE,DIVORCIADO/A,SOLTERO/A,VIUDO/A}
@attribute 'Lugar Nacimiento' numeric
@attribute 'Lugar Residencia' numeric
@attribute Sexo {F,M}
@attribute Edad numeric
@attribute 'Edad Primera Consulta' numeric
@attribute 'Diagnostico Topografico' {' 0,0','CA IN SITU MAMA','ENFERMEDAD DE HODGKIN','LEUCE
@attribute 'Estadio Cancer' numeric
@attribute 'Numero Operaciones' numeric
@attribute 'Numero Hospitalizaciones' numeric
@attribute Brecha numeric
@attribute 'Tratamiento Abandona' {Abandona,Continua}
@data
 0,SOLTERO/A, 43, 0,F, 36, 44,'TM DE LA MAMA', 100, 8, 5, 1,Abandona
 0,CASADO/A, 58, 0,M, 67, 67,'TM COLON', 100, 50, 27, 1,Continua
 50,CASADO/A, 15, 0,M, 61, 65,' 0,0', 100, 41, 8, 1,Continua
 0,SOLTERO/A, 0, 7,F, 25, 36,'TM ESTOMAGO', 100, 0, 14, 1,Continua
 0,CASADO/A, 0, 0,F, 37, 45,'TM DE LA MAMA', 33, 16, 8, 1,Continua
 0,CASADO/A, 18, 23,F, 46, 46,'TM ESTOMAGO', 0, 8, 8, 6,Continua
 0,CASADO/A, 28, 0,F, 53, 59,'LINFOMA NO HODGKIN FOLICULAR (NODULAR)', 66, 8, 33, 0,Continua
 0,SOLTERO/A, 15, 0,F, 44, 51,'TM DE LA MAMA', 66, 0, 13, 1,Continua
```

**Figura 5.1.** Estructura del archivo de tipo Arff para el proyecto. Imagen propia.

### 5.1.2 Propiedades de cada Conjunto de Datos

A continuación se mostrarán las variables que se tomarán en cuenta para el modelo, después de haber pasado por la etapa de pre procesamiento, y el respectivo tipo de dato para cada variable.

**Tabla 5.1**. Cuadro de variables del modelo.
Elaboración propia.

| Variables | Definición | Tipo de Dato |
|---|---|---|
| Condición Económica | Representa el seguro médico con el cual se atiende el paciente | Numérico |
| Estado civil | Estado determinado por las relaciones de familia | Cadena de caracteres |
| Lugar nacimiento | Distancia desde el lugar de nacimiento a Lima | Numérico |
| Lugar residencia | Distancia desde el lugar de residencia a Lima | Numérico |
| Sexo | Género del paciente | Cadena de caracteres |
| Edad | Edad del paciente | Numérico |
| Primera Consulta | Edad del paciente en su primera consulta | Numérico |
| Diagnostico topográfico | Órgano donde se encuentra el tumor | Cadena de caracteres |
| Estadio | Grado en el que se encuentra el cáncer | Numérico |
| Numero operaciones | Número de operaciones en el último año | Numérico |
| Numero hospitalizaciones | Número de hospitalizaciones en el último año | Numérico |
| Brecha | Brecha entre la fecha de diagnóstico y la fecha de inicio de tratamiento | Numérico |
| Tratamiento Abandona | Función Objetivo de abandono de tratamiento: Continua o Abandona | Cadena de caracteres |

### 5.1.3 Análisis de cantidades

Finalmente, es importante analizar si la cantidad de muestras por cada grupo de clasificación permitirá que se pueda realizar un modelo adecuado que otorgue resultados de alta precisión. Por tal motivo, a continuación se muestra el cuadro de cantidades de muestras por las clases que participaran en el modelo:

**Tabla 5.2**. Cuadro de cantidades por clase.
Elaboración propia.

| Clases | Cantidad | Porcentaje |
|---|---|---|
| **Continúa** | 3470 | 78.42% |
| **Abandona** | 955 | 21.58% |
| **Total** | **4425** | **100.00%** |

Las 4425 muestras finales, las cuales resultaron de la etapa de pre procesamiento, pertenecen a pacientes que empezaron un tratamiento de quimioterapia durante el 2013. A continuación se definirá cada clase de la función objetivo:

- **Continua:** Representa a los pacientes que continuaron con el tratamiento que se les indico para su enfermedad.
- **Abandona:** Representa a los pacientes que llevaron menos del 50% de sus sesiones programadas sin completar el tratamiento y no se tiene más registro para ellos sobre dicho tratamiento. En caso el paciente vuelva a su tratamiento después de un tiempo estimado por los miembros de la institución, este es considerado como un tratamiento distinto.

### 5.1.4 Conclusiones

Se ha realizado el análisis y definición del modelo de datos que se tomará en cuenta para el proceso iterativo de validación cruzada con cada algoritmo. La distribución para cada clase de la función objetivo o variable a clasificar se debe a que la mayoría de pacientes continúa sus tratamientos; sin embargo, dentro de la muestra, se ha podido obtener una cantidad considerable de abandono de tratamiento, lo cual contribuirá a una mejor decisión del modelo.

## 5.2 Resultado Esperado 7: Análisis de los algoritmos escogidos de Minería de Datos

Este resultado esperado consiste en realizar un análisis de los algoritmos que participaran dentro del proceso iterativo de Minería de Datos para determinar cuál de ellos se adecua mejor al modelo y a las condiciones actuales del conjunto de datos.

Además, se mencionará en la siguiente sección como se ha llevado a cabo su uso utilizando librerías de Minería de Datos existentes.

### 5.2.1 Análisis de Algoritmos

En esta sección se comentarán los cuatro algoritmos de clasificación que han sido tomados en cuenta para el modelo de acuerdo a la documentación revisada y a ciertas propiedades que son aceptables para el escenario que se pretende analizar. A continuación se presentará una breve definición de cada uno de ellos:

- **Árbol de decisión (Algoritmo J48-Implementación de C4.5):** Está basado en métodos lógicos donde se pueden diferenciar dos tipos de nodos en el árbol: una hoja representa una clase escogida y un nodo de decisión, la condición para un atributo. A partir de ello se realiza una búsqueda de la raíz hasta encontrar una hoja dónde se cumpla mejor la condición, y a partir de ello, clasifica una instancia o muestra [MEH 2011].

- **Support Vector Machine (Algoritmo SMO):** Está basado en el concepto de definir planos multidimensionales (hiperplanos) que permitan delimitar regiones complejas entre las diferentes clases que se pretenden clasificar. El algoritmo SMO es efectivo para realizar una clasificación no lineal [MEH 2011].

- **Redes Neuronales (Algoritmo MultilayerPerceptron):** Está basado en la idea de cómo opera un cerebro humano. Consiste de un número de nodos conectados por enlaces direccionales [MEH 2011]. En el caso del algoritmo MultilayerPerceptron, la red consta de tres elementos: una capa de entrada, una capa escondida que permite que los nodos extraigan, progresivamente, patrones complejos de las variables y una capa de salida dónde se observa la respuesta de clasificación, la cuál es más efectiva que el algoritmo de una sola capa [IAN 2011].

- **k-Vecinos más Cercanos (Algoritmo IBk):** Está basado en un método de clasificación por instancias de manera local; es decir, escogerá la clase de los k vecinos que se encuentren más cerca a la muestra que se desea clasificar [MEH 2011]. El algoritmo IBk permite seleccionar, automáticamente, el valor más apropiado para k y brindar un peso para cada variable [IAN 2011].

A continuación se presentará un cuadro comparativo de las principales ventajas y desventajas que presenta cada uno de los cuatro algoritmos que se utilizarán en el modelo [MEH 2011]:

**Tabla 5.3**. Ventajas y Desventajas de los algoritmos elegidos.
Elaboración propia.

| | **Algoritmo** | **Ventajas** | **Desventajas** |
|---|---|---|---|
| **Arboles de Decisión** | J48 | -Efectivo para conjunto de datos de pocas clases<br>-Se puede obtener el diagrama del árbol de decisión | -Define áreas muy simples (lineales) |
| **Redes Neuronales** | Multiperceptron | -Permite encontrar patrones en conjuntos de datos con mucha distorsión.<br>-Permite lidiar con la variación de las condiciones de la data<br>-Facilita la limitación de grupos en zonas más complejas (no lineal) | -Su tiempo de ejecución es muy alto |
| **SVM** | SMO (no lineal) | -Permite definir regiones de forma compleja (no lineal), aumentando la dimensionalidad<br>-Posee un buen rendimiento para pequeños conjuntos de entrenamiento | -Presenta un límite de patrones del conjunto de entrenamiento |
| **Knn** | IBk | -Su tiempo de ejecución es rápido para un conjunto de datos pequeño.<br>-Busca el valor más apropiado para k | -Requiere de una buena normalización<br>-Menor efectividad cuando los datos son valores nominales. |

### 5.2.2 Herramienta Weka

La herramienta Weka, como se explicó en la sección de Herramientas del capítulo 2, es una plataforma libre que ofrece una colección de máquinas de aprendizaje para

desarrollar tareas de Minería de Datos, entre ellas, la clasificación. Los algoritmos que se han mencionado en la sección de análisis de algoritmos se encuentran ya implementados dentro de las librerías de Weka. Además, Weka ofrece las siguientes opciones que serán de vital importancia para la evaluación de la precisión de los algoritmos que se presentará en el siguiente capítulo [WEK 2014]:

- Validación cruzada y configuración del parámetro k
- Resultados de precisión
- Matriz de Confusión
- Curvas ROC

### 5.2.3 Conclusiones

En esta sección se ha presentado el análisis de los algoritmos, el cual es muy importante para poder entender de manera general el funcionamiento de cada uno y las razones por las cuales fue conveniente tomarlo en cuenta para el modelo de Minería de Datos. Además, se ha podido relacionar cada técnica de clasificación con los algoritmos que ofrece la plataforma Weka, en la cual se pueden obtener los algoritmos ya implementados y diversas funcionalidades que apoyarán el proceso de elección del algoritmo con mayor precisión.

## 5.3 Resultado Esperado 8: Evaluación de la precisión de los algoritmos y elección del algoritmo más adecuado

Se ha evaluado cada algoritmo a través del método de validación cruzada para conseguir el algoritmo que mejor se adecúe a las condiciones del modelo, con el fin de que sea utilizado dentro del proceso automatizado final. Dentro del conjunto de métodos y procedimientos que se han realizado en la etapa de pre procesamiento, es importante identificar cuál es la mejora en el desempeño del clasificador que la normalización y el análisis de valores anómalos han aportado al modelo final y cuáles son los parámetros ideales para que se puedan lograr los objetivos de precisión del modelo.

### 5.3.1 Evaluación de los algoritmos por el método de validación cruzada

Como parte del proceso de Minería de Datos, es importante evaluar los algoritmos escogidos en un conjunto de iteraciones del modelo para determinar quien en la mayoría de los casos es el algoritmo que nos proporciona mejores resultados. El método de validación cruzada permite dividir el conjunto de datos en k subconjuntos y utilizar cada subconjunto como prueba para cada una de las k iteraciones, resultando una precisión promedio y otros datos que provee Weka [IAN 2011]. Por esta razón, este método es ideal al momento de comparar procedimientos o algoritmos de Minería de Datos.

De acuerdo a documentación revisada, se suele usar valores de k=10 para una población como la que se está trabajando, e incluso, es la que aparece por defecto en la herramienta Weka. Las pruebas realizadas mencionan que es el valor más aceptable cuando se comparan los algoritmos por el método de validación cruzada [IAN 2011]:

**Tabla 5.4**. Resultados del método de validación cruzada.
Elaboración propia.

| | **k=10** |
|---|---|
| **Árboles de decisión** | 89.98% |
| **Redes neuronales** | 84.15% |
| **SVM** | 81.62% |
| **Knn** | 83.16% |
| **Promedio** | **84.73%** |

Como se mencionó en el resultado, obtener una precisión mayor a 60% le otorga un alto grado de confiabilidad al modelo, lo cual se está logrando para todos los algoritmos.

### 5.3.2 Elección del algoritmo más eficaz

Como se puede apreciar en la tabla 5.4, el algoritmo que en promedio otorga mejores resultados de precisión es el J48, perteneciente a los árboles de decisión. Sin embargo, en conjuntos de datos donde la distribución de clases no es equitativa (solo el 21.58% pertenece a la clase abandono de tratamiento), se requieren de otros

métodos que permitan medir de manera más apropiada el rendimiento del modelo. Por esta razón, se han utilizado los siguientes dos métodos [MEH 2011]:

- **Matriz de Confusión:** La matriz de confusión contiene el número de aciertos y el número de instancias mal clasificadas (la clase que escogió) para cada clase. Este método permite determinar la precisión de clasificación para cada clase y en que situaciones se confunde o comete el error de escoger determinada clase.

- **Curvas Roc:** Es un gráfico que representa la relación de sensibilidad del modelo: razón entre el ratio de verdaderos positivos (porcentaje de acierto de la clase abandona tratamiento) y el ratio falsos positivos (porcentaje de acierto de la clase continua tratamiento). Estos ratios provienen de la matriz de confusión. Mientras más cerca se encuentre a 1 el resultado de las curvas ROC, será considerado un modelo más óptimo.

Árboles

| a | B | Total | <-- classified as |
|---|---|---|---|
| 670 | 285 | 955 | a=Abandona |
| 158 | 3312 | 3470 | b=Continua |

Redes Neuronales

| a | B | Total | <-- classified as |
|---|---|---|---|
| 464 | 491 | 955 | a=Abandona |
| 210 | 3260 | 3470 | b=Continua |

SVM

| a | B | Total | <-- classified as |
|---|---|---|---|
| 224 | 731 | 955 | a=Abandona |
| 82 | 3388 | 3470 | b=Continua |

KNN

| a | B | Total | <-- classified as |
|---|---|---|---|
| 482 | 473 | 955 | a=Abandona |
| 272 | 3198 | 3470 | b=Continua |

**Figura 5.2**. Matriz de Confusión para cada algoritmo.
Elaboración propia.

La figura 5.2 muestra las matrices de confusión que se han obtenido para cada algoritmo. Las columnas representan como fue clasificada cada instancia, mientras que las filas representan la clase real a la que pertenecía cada instancia. Un algoritmo con buen rendimiento debería tener la mayor cantidad de instancias en la diagonal (celdas sombreadas), lo cual representa la cantidad de instancias que han sido clasificadas bien. Como se puede observar, los algoritmos de árboles y kNN son los que mejores resultados nos ofrecen; sin embargo, el objetivo del proyecto está enfocado en determinar los pacientes con tendencia a abandonar el tratamiento, con el fin de que se puedan plantear estrategias de retenciones bien direccionadas para dichos pacientes. Por esta razón, el algoritmo más ideal de los dos para el proyecto es el que tiene un mayor acierto en la clase abandono de tratamiento.

**Tabla 5.5**. Sensibilidad y Especificidad para cada algoritmo.
Elaboración propia.

| | Sensibilidad (Acierto Abandona) | Especificidad (Acierto Continua) |
|---|---|---|
| **Árboles de decisión** | 70.15% | 95.44% |
| **Redes neuronales** | 48.58% | 93.94% |
| **SVM** | 23.45% | 97.63% |
| **kNN** | 50.47% | 92.16% |

En la tabla 5.5 se puede observar los porcentajes de acierto, tanto para la clase abandona como continua, obtenidos por cada algoritmo. A partir de ello, el algoritmo más adecuado es el J48.

Como siguiente paso se muestra el resultado de las curvas ROC, el cual, finalmente, indicará el algoritmo más adecuado para el modelo, basándose en la distribución de instancias para cada clase:

**Tabla 5.6**. Curvas de ROC para cada algoritmo.
Elaboración propia.

| | Curvas ROC |
|---|---|
| **Árboles de decisión** | 0.877 |
| **Redes neuronales** | 0.796 |
| **SVM** | 0.605 |
| **Knn** | 0.845 |

Como se muestra en la tabla 5.6, el algoritmo J48 (árboles de decisión) e ibK (kNN) son los que tienen resultados de curva de ROC más cercanos a 1; es decir, son considerados los más óptimos para el modelo. Sin embargo, el algoritmo de árboles de decisión es el que mejor resultados de precisión nos ofrece, sobre todo para la clase de Abandona. Por tal motivo, después de comparar los algoritmos por la matriz de confusión y las curvas ROC se ha determinado que el algoritmo de árboles de decisión será utilizado en el proceso automatizado final.

### 5.3.3 Evaluación de desempeño en la normalización

La importancia de realizar un procedimiento de normalización de datos radica, principalmente, en la mejora de la precisión para ciertos algoritmos y en una reducción significativa del tiempo de construcción del modelo por cada algoritmo. A continuación

se muestran los resultados obtenidos antes y después de la normalización para los cuatro algoritmos:

**Tabla 5.7**. Resultados de desempeño-Normalización.
Elaboración propia.

| | Pre-Normalización | | Normalizado | |
|---|---|---|---|---|
| | **Tiempo (seg)** | **Precisión** | **Tiempo (seg)** | **Precisión** |
| **Árboles de decisión** | 1.14 | 87.70% | 0.36 | 89.98% |
| **Redes neuronales** | +(10000) | - | 419.08 | 84.15% |
| **SVM** | 636.91 | 82.24% | 21.53 | 81.62% |
| **kNN** | 0.01 | 82.76% | 0.01 | 83.16% |

Como se puede apreciar en la tabla 5.7, se ha conseguido un aumento en la precisión en general, siendo el más significativo en el árbol de decisión. Esto se debe a que los árboles de decisión son más efectivos cuando los valores de las variables no son tan dispersos. El algoritmo de árboles de decisión también ha tenido una mejora en tiempo, lo cual es primordial ya que es el algoritmo que se utilizará para el proceso. Con respecto al tiempo, en todos los algoritmos se puede ver una mejora sustancial, sobretodo en el algoritmo de redes neuronales. El factor tiempo es de vital importancia ya que al momento de realizarse el proceso automatizado, se tomarán en cuenta para el modelo todas las muestras de los pacientes que hayan recibido tratamiento con un mínimo de tres años de anterioridad, por lo cual, serán aproximadamente 20 mil instancias que podrían hacer que el modelo tardara mucho tiempo.

Cabe resaltar que el proceso de normalización también tiene influencia sobre el algoritmo de detección y eliminación de valores anómalos debido a que permitirá que el cálculo de las distancias se haga de una manera más rápida y efectiva, sobre todo al transformar y agrupar los valores de tipo cadena de caracteres.

### 5.3.4 Evaluación del desempeño del algoritmo de análisis de valores anómalos

La importancia del algoritmo de análisis de valores anómalos radica en poder identificar las muestras que tengan un comportamiento diferente al de su clase (por lo cual podrían variar de forma negativa los patrones para cada clase) y eliminarlas del modelo final. Para este punto se comprobará el aumento de la precisión para el algoritmo J48 (árboles de decisión), de acuerdo a la configuración de parámetros k y umbral que se haya escogido. Los valores que comúnmente se escogen para k son 3 o 5, lo cual depende también del tamaño del conjunto de datos [MEH 2011]. A

continuación se muestran los resultados para el algoritmo de detección de valores anómalos:

**Tabla 5.8.** Resultados de desempeño-Análisis de Outliers.
Elaboración propia.

| 89.98% | **k=3** | **k=5** | **k=10** |
|---|---|---|---|
| **Umbral=2** | 93.01% | 92.45% | 90.33% |
| **Umbral=3** | 89.72% | 92.84% | 91.40% |
| **Umbral=4** | 84.63% | 92.25% | 91.64% |

La tabla 5.8 muestra la precisión para cada configuración de parámetros k y umbral, donde, finalmente, se puede identificar que los valores ideales para cada parámetro, los cuales se han tomado en cuenta en el algoritmo, son:

- k= 3
- Umbral= 2

Estos valores aparecerán por defecto en el proceso automatizado; sin embargo, podrán ser probados y calibrados dependiendo de que tanto mejoren el desempeño para el modelo actualizado.

### 5.3.5 Conclusiones

Se ha realizado la evaluación de los algoritmos por el método de validación cruzada, comparando los algoritmos con el valor de k como 10. Una vez definido este parámetro, se han comparado los resultados por los dos métodos de medición, los cuales han guiado a la elección del algoritmo de árboles de decisión (J48) para ser utilizado en el proceso automatizado de Minería de Datos.

Finalmente, se ha comprobado la utilidad de los métodos de normalización y análisis de valores anómalos en la mejora de la precisión y los tiempos de ejecución del modelo. En el caso del análisis de valores anómalos, se puede apreciar que dependiendo de ciertas configuraciones de los parámetros, se puede obtener una mayor precisión en cada situación. Esta prueba solo se ha realizado con el algoritmo escogido para determinar cuál es la forma en que mejor se adecúen los parámetros al proceso automatizado.

# Capítulo 6: Automatización del proceso de Minería de Datos

En este capítulo se explicarán los resultados obtenidos para el cuarto y último objetivo específico, el cual consiste en automatizar el proceso de Minería de Datos para la carga de nuevo pacientes, con tal de que el modelo pueda adaptarse a las condiciones futuras del tratamiento de quimioterapia. Para este objetivo se realizará el ejecutable que contiene el proceso de actualización del modelo y el reporte con los resultados para cada paciente.

## 5.4 Resultado Esperado 9: Programa ejecutable para la actualización del proceso de Minería de Datos

Una vez que se ha escogido el algoritmo de clasificación, es de vital importancia tener en cuenta que cada método pueda ser realizado de forma automatizada para poder, finalmente, predecir que pacientes tienen tendencia a abandonar el tratamiento, de tal manera que no requiera una gran operación por parte de los miembros de la institución. Además, se requiere que el modelo pueda adaptarse a las nuevas condiciones del tratamiento, por lo cual el modelo se actualizará de acuerdo a cómo se programen las ejecuciones del proceso automatizado.

Para la elaboración del proceso automatizado se ha utilizado la herramienta Pentaho, al igual que en la carga del DataMart. Esto permitirá integrar ambos procesos con el fin de que puedan actualizarse periódicamente.

### 5.4.1 Proceso Automatizado de Minería de Datos

El proceso automatizado contará con los métodos de normalización, análisis de valores anómalos y predicción de los pacientes con tendencia a abandonar el tratamiento. Del proceso de normalización de datos se tendrán dos conjuntos de datos como salidas: el primero formará parte del modelo y el segundo contendrá las características de los pacientes que se desea clasificar. Cabe resaltar que, con respecto al método de análisis de valores anómalos, solo se realizará para la salida que representa el modelo. Como resultado final, se generarán los reportes para cada departamento de atención con la salida que genere el algoritmo de árboles de decisión

y se enviará la confirmación de que el proceso ha sido actualizado con éxito a cada uno de los interesados.

A continuación, en la figura 6.1, se puede observar el flujo general que seguirá el proceso automatizado de Minería de Datos

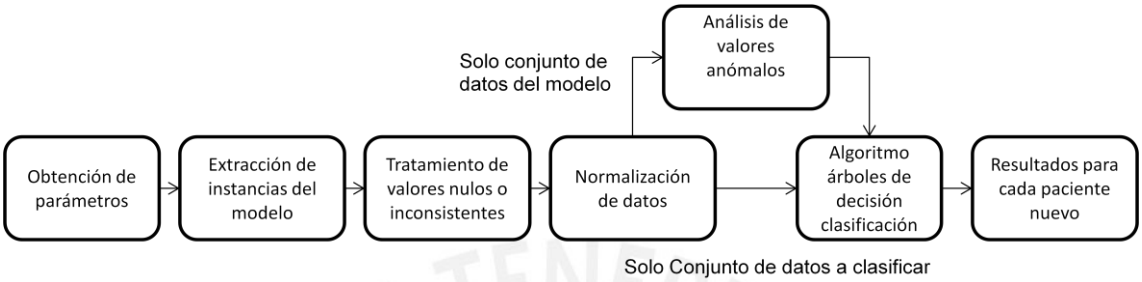


**Figura 6.1**. Flujo general del proceso automatizado de Minería de Datos. Elaboración propia.

En el anexo 10.5 se puede observar el mapa de los principales procesos involucrados en el proceso automatizado de Minería de Datos.

En la figura 6.2 se muestra como ejemplo el proceso que se realizará para el análisis de cada nueva instancia utilizando los algoritmos de la librería Weka, lo cual finalmente proveerá los reportes de abandono. Como se puede apreciar en la figura 6.2, en primera instancia, se verifica que exista el archivo y luego se moverá dicho archivo a la carpeta adecuada. Finalmente, se realizará la ejecución del archivo ".jar" que contiene la aplicación con el algoritmo J48 (árboles de decisión) y la generación de los distintos tipos de reportes.

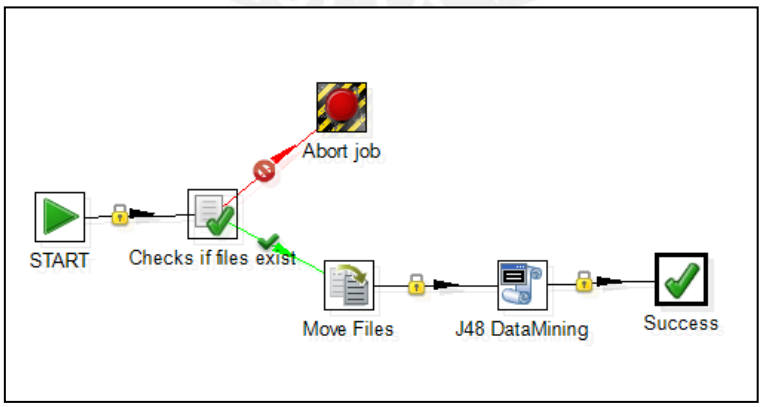


**Figura 6.2**. Proceso automatizado para ejecutar el algoritmo J48. Elaboración propia.

### 5.4.2 Ventana de Configuración de Parámetros

Como parte del proceso automatizado es importante tener en cuenta los parámetros que cada método requiere y que puedan ser fácilmente configurables por los miembros de la entidad. Por tal motivo, se ha realizado una aplicación que permita configurar estos valores de acuerdo a como crea más conveniente la institución. Entre las funcionalidades que se han previsto se encuentran:

- Actualización de los parámetros de detección de valores anómalos.
- Pruebas de los parámetros de valores anómalos con el modelo actual.
- Actualización de parámetros del modelo (tamaño del modelo en meses, punto de quiebre, entre otros).
- Actualización de los parámetros de los reportes de salida.

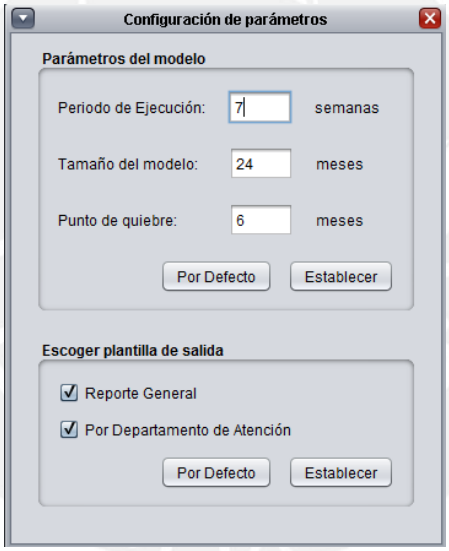


**Figura 6.3**. Ventana de Configuración de parámetros del modelo.
Elaboración propia.

En la figura 6.3, se puede observar la ventana que contiene los parámetros principales que determinarán el periodo de ejecución, el tamaño del modelo y las fechas que se tomarán en cuenta. En este caso, por ejemplo, el modelo contemplará los tratamientos iniciados en las fechas 12/08/2013 al 12/08/2014, tomando como fecha actual el 12/10/2014. Con respecto al periodo de ejecución, este valor ha sido establecido como semanal con los miembros del área de sistemas de la institución. Lo ideal es que se actualice en el mismo periodo que el DataMart. Además, se pueden establecer los tipos de plantillas de los reportes que serán resultado del proceso automatizado.

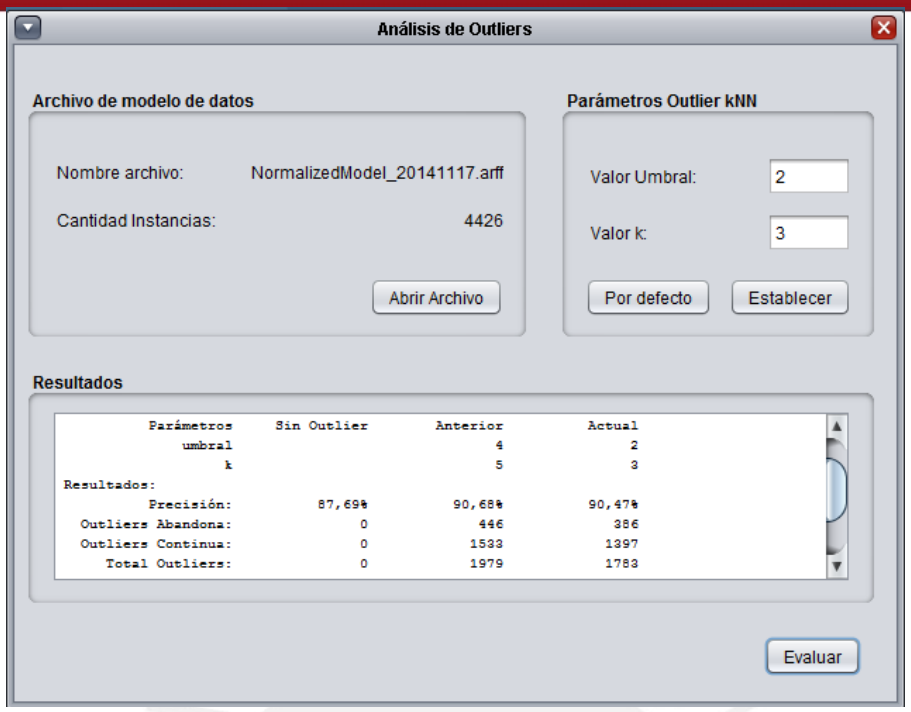


**Figura 6.4**. Ventana de configuración y prueba de parámetros del algoritmo de detección de valores anómalos. Elaboración propia.

En la figura 6.4, se puede observar la ventana de calibración de los parámetros “umbral” y “k” para el algoritmo de detección de valores anómalos. Esta ventana permitirá realizar pruebas de precisión con el algoritmo escogido para determinar los parámetros más adecuados para cada ejecución debido a que el modelo se actualiza periódicamente, lo cual puede hacer que una nueva configuración de parámetros sea más eficiente. Para esto, se realizará una comparación manual con los resultados que se obtengan entre los parámetros establecidos y los que se coloquen en los campos respectivos. Finalmente, dentro de esta misma ventana se pueden establecer o actualizar los valores que utilizará el modelo.

### 5.4.3 Conclusiones

Se ha realizado la automatización de los procesos relacionados con el pre procesamiento del modelo y la predicción para los nuevos pacientes. Además, se han realizado las ventanas que permitirán configurar adecuadamente los parámetros para todo el proceso en general y también poder probar los parámetros de valores anómalos de acuerdo a los resultados del algoritmo escogido. Esto es importante

debido a que permitirá que los miembros de la institución puedan adaptar el proceso de Minería de Datos a las condiciones actuales del tratamiento.

Finalmente, el proceso automatizado debe proveer a los interesados el reporte o informe con los pacientes con tendencia a abandonar el tratamiento. Esto forma parte de la última etapa de la metodología, la cual consiste en la interpretación y conclusiones y se abordará en el siguiente resultado esperado.

## 5.5 Resultado Esperado 10: Reporte con los resultados para cada paciente

Como etapa final de un proceso de Minería de Datos es importante que los resultados del modelo puedan ser correctamente interpretados y resumidos, con el fin de que se pueda ayudar a los miembros de la organización en la toma de decisiones. Los administradores de la entidad de salud no esperan grandes cantidades de información, sino más bien, una información resumida que puedan entender para una exitosa toma de decisiones [MEH 2011].

Por esta razón, se ha realizado la creación de reportes con la información de los pacientes que presentan tendencia a abandonar un tratamiento de quimioterapia. Estos reportes se proporcionarán periódicamente de acuerdo a como sea ejecutado el modelo con la información para los nuevos tratamientos. Entre los reportes que se han tomado en cuenta se encuentran:

- Reporte por departamento de atención: Se generará un reporte por cada departamento de atención, donde se mostrará la lista de pacientes cuya predicción resulto en la clase “abandona”. Se tendrá en cuenta variables principales del paciente como sexo, edad, lugar de residencia, condición socioeconómica, diagnóstico topográfico, estadío del cáncer y un número telefónico, en caso se desee contactar con el paciente.

- Reporte General: Se generará un solo reporte que mostrará una lista de la cantidad de pacientes que abandonaron y continuaron sus tratamientos agrupada por departamento de atención, lo cual podría servir para seguir midiendo la precisión del modelo en el futuro. Finalmente, se mostrará un gráfico que presente esta información resumida. El objetivo de este reporte es

que pueda servir de apoyo a los niveles superiores encargados de tomar las decisiones.

**Reporte de Pacientes**

Abandono de Tratamientos

INEN

**Tratamiento:** Quimioterapia

**Departamento de atención:** MAMAS Y TEJIDOS

| Código Paciente | Sexo | Edad | Lugar de Residencia | Condición Socioeconómica | Diagnóstico Topográfico del Cáncer | Estadio Cáncer | Teléfono |
|---|---|---|---|---|---|---|---|
| 13 | M | 12 | LIMA | S.I.S (E) | LEUCEMIA LINFOIDE | III | 984747474 |
| 23 | M | 63 | LIMA | S.I.S (E) | TM ESTOMAGO | II | 984747474 |
| 886 | M | 9 | HUANUCO | S.I.S (E) | LEUCEMIA MIELOIDE | II | 984747474 |
| 54 | M | 64 | LIMA | S.I.S (E) | TM CUELLO UTERINO | II | 984747474 |
| 898 | F | 77 | LIMA | S.I.S (E) | TM DE LA MAMA | IV | 984747474 |
| 42 | F | 45 | LIMA | S.I.S (E) | 0,0 | IV | 984747474 |
| 1496 | F | 48 | LIMA | S.I.S (E) | TM DE LA MAMA | II | 984747474 |
| 329 | F | 50 | LIMA | S.I.S (E) | TM DE LA MAMA | III | 984747474 |

**Figura 6.5**. Ejemplo de reporte de abandono de tratamientos por departamento de atención. Elaboración propia.

En la figura 6.5 se puede observar un fragmento de salida del reporte de abandono de tratamiento de quimioterapia por departamento de atención en formato pdf el cual será enviado a los interesados por departamento de atención.

### 5.5.1 Conclusiones

Se han realizado los distintos reportes con la información de salida del modelo de Minería de Datos y la respectiva interpretación de los resultados, con lo cual se llega al final del proceso de Minería de Datos. Dentro de los reportes, se ha procurado que sean entendibles y manejables para los miembros de la organización que lo utilizarán en la toma de decisiones.

# 6 Discusión de Resultados

En esta sección se presentará la discusión de los resultados que se han obtenido a lo largo del proyecto. Para cada objetivo específico se mostrará el problema que se planteó solucionar y los logros que se han obtenido para corroborar que se han cumplido las expectativas de cada uno de ellos. Finalmente, se han presentado las propuestas con respecto a cada objetivo, con el fin de que se puedan mencionar recomendaciones que guíen a trabajos futuros tomando como base los resultados de este proyecto.

## 6.1 Objetivo específico 1

**Problema:** El problema que se ha encontrado en la institución es que la información que se requiere no está integrada y organizada de la forma más adecuada, lo cual puede complicar las labores de extracción de información.

**Solución:** Como solución se propuso la creación de un modelo de DataMart de pacientes y los procesos ETL que permitan que la información pueda cargarse de forma automatizada.

**Logros:** Se realizaron reuniones con los miembros de la institución con el propósito de diseñar el modelo más adecuado para la institución, basándose en la base de datos con la que ellos operan diariamente y la problemática del proyecto de académico. Finalmente, se diseñaron los procesos ETL para que la información pueda cargarse de forma automatizada al DataMart y sean fáciles de mantener por los miembros de la institución.

**Propuesta:** La creación de un DataMart especializado en el tema de la problemática puede dar lugar a que se vea con buenos ojos las labores de integración de la información. Como propuesta, se podría realizar la creación de un Data Warehouse, el cual pueda ayudar a la toma de decisiones de la organización en general.

## 6.2 Objetivo Específico 2

**Problema:** El problema que se ha encontrado es que, en sistemas con grandes cantidades de información, pueden existir instancias con variaciones y anomalías en sus características que podrían distorsionar la formación de patrones del modelo y reducir su desempeño al momento de la evaluación.

**Solución:** Como solución se ha planteado enfocarse en la etapa de pre procesamiento con el fin de afinar el modelo de datos. Dentro de esta etapa se ha contemplado el análisis de los valores nulos, un método de normalización y, finalmente, el algoritmo de detección y eliminación de Valores anómalos.

**Logros:** Se ha realizado el análisis para decidir que variables cuentan con suficientes datos consistentes para ser tomados en cuenta en el modelo. Este proceso también se ha tomado en cuenta en las instancias que integrarán el modelo. Con respecto a la normalización, se ha demostrado que existe una mejora, tanto en tiempo como en precisión, lo cual brinda una mayor confiabilidad a los resultados del modelo. Finalmente, el algoritmo kNN, para la detección y eliminación de valores anómalos, ha permitido que el modelo pueda generar patrones mejor definidos, lo cual también ha aumentado el desempeño al momento de la predicción para el algoritmo escogido.

**Propuesta:** Como propuesta se plantea el uso de un algoritmo kNN optimizado u otra técnica de detección de valores anómalos (clusterización, algoritmos estadísticos, entre otros) que puedan aumentar el desempeño del modelo. Como otra opción se puede implementar una técnica para determinar la influencia de cada variable y que el modelo pueda tener en cuenta solo las que sean más significativas.

## 6.3 Objetivo Específico 3

**Problema:** El problema que se ha encontrado es que la entidad de salud pública no posee los mecanismos para determinar la continuidad de los pacientes con respecto a su tratamiento de forma acertada.

**Solución:** Como solución se ha planteado la evaluación de un conjunto de algoritmos de clasificación, con el fin de obtener el que mejor se adecúe al escenario actual.

**Logros:** Se ha logrado obtener un conjunto de datos con variables bien definidas que permitirá construir un modelo confiable. Esto se demuestra en los resultados obtenidos con una precisión de 90%, aproximadamente, sobre el 60% que se planteó inicialmente como aceptable. Como resultado de la evaluación de los algoritmos se ha escogido el algoritmo J48 (árboles de decisión) utilizando como métricas los métodos de matriz de Confusión y curvas de Roc. Dentro de estos métodos se ha podido apreciar que el algoritmo permite enfocarse en el objetivo del proyecto que es predecir el grupo de pacientes con tendencia a abandonar el tratamiento con mayor precisión. Finalmente, también se realizaron las evaluaciones de los métodos de normalización y análisis de valores anómalos, con el fin de verificar su importancia en el modelo y establecer los parámetros más adecuados.

**Propuesta:** Como propuesta se plantea la evaluación de los algoritmos tomando en cuenta diferentes variables para el modelo y que tanto podría influir eso en una mejor decisión. Además, se plantea implementar un algoritmo optimizado, diferente al de los existentes en las librerías de Weka, que permita poder determinar también la causa principal que lleva a los pacientes a abandonar sus tratamientos. Esto podría estar acompañado de nuevas variables que puedan jugar un papel importante en la decisión d abandono. Se podría utilizar el algoritmo de árboles de decisión para identificar las reglas y así obtener las variables más influyentes. Finalmente, se podría tomar en cuenta como clase objetivo el regreso de un paciente a su tratamiento.

## 6.4 Objetivo Específico 4

**Problema:** El problema que se ha encontrado es que debido a la gran cantidad de procesos existentes en la institución, sería muy difícil dar seguimiento a cada proceso de Minería de Datos y a la entrega de los resultados.

**Solución:** Como solución, se ha planteado automatizar el proceso de Minería de Datos desde su etapa de extracción hasta la etapa de interpretación y entrega de resultados.

**Logros:** Se ha conseguido automatizar las etapas mencionadas en el proceso de Minería de Datos, brindando los resultados de manera oportuna a los interesados del proceso. Este proceso tiene una herramienta de configuraciones que permitirá

modificar los parámetros necesarios para dar mantenimiento al modelo y lograr mejorar el análisis de valores anómalos dependiendo del modelo actualizado. Finalmente, se ha procedido a entregar los resultados en un formato entendible para los miembros de la organización y con información de utilidad

**Propuesta:** Como una propuesta se plantea que se pueda utilizar el método automatizado para generar otro tipo de conocimiento como predecir las causas principales de abandono, determinar el tiempo de duración de un tratamiento, entre otros.

# 7 Conclusiones

Durante cada sección del proyecto académico de fin de carrera, se ha buscado enfatizar la importancia de seguir cada paso perteneciente al proceso de Minería de Datos. La metodología no solo involucra la prueba de los algoritmos, sino es un proceso que inicia en el establecimiento del problema, de vital importancia para no desviarse del objetivo final, la correcta extracción y procesamiento de los datos, hasta el despliegue e interpretación de resultados a los miembros de la institución. La importancia de que este proceso sea automatizado radica en liberar de tareas adicionales a los miembros de la institución, así como brindar información oportuna y en los plazos establecidos a cada uno de ellos.

La motivación para realizar este proyecto es poder brindar una herramienta confiable que permita apoyar la toma de decisiones de una forma personalizada para los pacientes que puedan abandonar los tratamientos de quimioterapia. La generación de conocimiento en una entidad de salud pública utilizando un proceso de Minería de Datos proporciona una base precisa para que se puedan apoyar las estrategias con respecto a los diferentes problemas.

Con respecto al primer objetivo específico, se ha podido identificar la importancia de tener una información bien estructurada para facilitar su entendimiento y consistencia, más aún ya que la institución no cuenta con un DataWarehouse para el análisis de dicha información. Por esta razón, se realizaron una serie de reuniones con los miembros de la institución para poder comprender el entorno donde se va a implementar la solución. A partir de esto, se realizó el modelo de DataMart y el proceso de carga automatizada, el cual se ha adaptado a las necesidades, tanto del problema como de la institución. El modelo de DataMart, permitirá que en un futuro se pueda generar conocimiento para otros tipos de problemáticas.

Con respecto al segundo objetivo, el cual corresponde a los métodos para el pre procesamiento del conjunto de datos crudo, se ha podido confirmar su importancia en la mejora del desempeño de los resultados, tanto en tiempo como en precisión. Los métodos utilizados fueron el tratamiento de valores nulos, normalización y análisis de valores anómalos. Estos métodos han requerido de la participación de los miembros de la institución para poder determinar las variables más adecuadas a reemplazar para el escenario de los tratamientos de quimioterapia. Finalmente, el identificar las

instancias anómalas del modelo ha permitido una mayor definición de los patrones que se formarán, lo cual ha guiado a una mayor precisión de clasificación.

Con respecto al tercer objetivo, se realizó la definición de las variables que se tomaron en cuenta para el modelo habiendo realizado un análisis previo con los miembros de la institución sobre cuáles de ellas serían las más influyentes. Además, se realizó el análisis de los cuatro algoritmos que participaron en la evaluación para identificar la razón por las cuales se han escogido. En la etapa de evaluación, se pudo determinar la importancia de cada uno de los métodos de normalización utilizados y, finalmente, evaluar el algoritmo para escoger el más apropiado. El algoritmo que mejor resultado proporcionó, de acuerdo al análisis de la matriz de confusión y las curvas ROC, fue J48 (árboles de decisión); por tal motivo, este algoritmo ha sido tomado en cuenta para predecir la tendencia de abandono de los tratamientos en el proceso automatizado de Minería de Datos.

Con respecto al cuarto y último objetivo, se ha procedido con la automatización del proceso de Minería de Datos, con el objetivo de que se pueda generar conocimiento oportunamente a los interesados en el tema de los tratamientos. Como primer paso se realizó una ventana para que los encargados del mantenimiento puedan configurar los parámetros del modelo y del algoritmo de valores anómalos y, finalmente, probar el desempeño en cada escenario del modelo actualizado. El proceso automatizado se ha implementado para tomar en cuenta los métodos de pre procesamiento especificados y el algoritmo que permitirá predecir el comportamiento para los nuevos paciente. Como resultado final se han diseñado los reportes que se generarán por departamentos de atención y de manera general. Todos estos resultados que se han tomado en cuenta permitirán que el proceso de Minería de Datos sea sencillo de realizar y mantener para los miembros de la institución.

Finalmente, se ha culminado el proyecto de académico de fin de carrera entendiendo la importancia de convertir la información en conocimiento para apoyar la toma de decisiones. Además se ha podido identificar como influye tomar en cuenta cada etapa del proceso de Minería de Datos par que la solución a la problemática tenga una mayor efectividad. Cabe resaltar que la metodología planteada, no solo debe permitir resolver esta problemática, sino también servir como fuente para generar otro tipo de conocimiento que apoye a la institución

# 8 Referencias bibliográficas


[HOR 2007] Hornick, M.F., et al., Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for architecture, design, and implementation. 2007: Morgan Kaufmann Publishers Inc.

[OLA 2014] Ola, O. and K. Sedig, The challenge of big data in public health: an opportunity for visual analytics. Online J Public Health Inform, 2014.

[MIN1 2011] Perú, Ministerio de Salud. Lima: MINSA; 2011. http://www.minsa.gob.pe/ogei/conferenciaops/index.asp. Último acceso: abril 2014.

[INS 2012] Perú, Instituto Nacional de Salud. Lima: INS; 2012. http://www.peru.gob.pe/docs/PLANES/10032/PLAN_10032_PEI_2011_-_2015_Informe_General_2012_2013.pdf. Último acceso: abril 2014.

[MIN2 2011] Perú, Ministerio de Salud. Lima: MINSA; 2011. http://www.minsa.gob.pe/portada/Especiales/2011/respiravida/archivos/Ayuda_memoria_Lanzamiento_TB.pdf. Último acceso: abril 2014.

[INE 2006] Perú, Instituto Nacional de Enfermedades Neoplásicas. Lima: INEN; 2006. http://www.inen.sld.pe/portal/documentos/pdf/normas_tecnicas/2006/25052012_PLAN_NAC_PREV_CONTROL_CA.pdf. Último acceso: abril 2014.

[INS 2011] Perú, Instituto Nacional de Salud. Lima: INS; 2011. http://www.ins.gob.pe/repositorioaps/0/4/jer/evidencias/Nota%20t%C3%A9cnica-10_Di%C3%A1logo%20deliberativo%20Intervenciones%20dirigidas%20a%20disminuir%20el%20abandono%20al%20tratamiento%20antituberculosos.pdf. Último acceso: abril 2014.

[MEH 2011] Kantardzic, M., Data mining: concepts, models, methods, and algorithms. 2011: John Wiley & Sons.

[WEI 2010] Weiss, G.M. and B.D. Davison, Data Mining, in HANDBOOK OF TECHNOLOGY MANAGEMENT, H. BIDGOLI (ED.). 2010, John Wiley and Sons. pp. 1-17.

[FAY 1996] Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, M.F. Usama, et al., Editors. 1996, American Association for Artificial Intelligence. pp. 37-54.

[RAE 2014] España, Real Academia Española. Madrid: RAE; 2014 http://www.rae.es/. Último acceso: abril 2014

[MIN 2005] Perú, Ministerio de Salud. Lima: MINSA; 2005. http://www.minsa.gob.pe/hama/Informaci%C3%B3n_Hma/Estadistica/Norma%20HC%20V02.pdf. Último acceso: mayo 2014.

[PAH 2003] Estados Unidos, Organización Panamericana de la Salud: PAHO; 2003. http://ais.paho.org/classifications/Chapters/pdf/Volume2.pdf. Último acceso: mayo 2014.

[UNF 2012] Perú, Fondo de Población de las Naciones Unidas-Perú: UNFPA; 2012. http://www.unfpa.org.pe/Legislacion/PDF/20120717-MINSA-NT-Atencion-Adulto-VIH.pdf. Último acceso: mayo 2014.

[ROB 2013] Robert Porter, Mark. An Analysis of Treatment Retention and Attrition in an Australian Therapeutic Community for Substance Abuse Treatment. 2013. Título para Doctor de Psicología. Australia: Edith Cowan University.

[HAC 2013] Hachesu, P.R., et al., Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res, 2013. pp. 121-129.

[ZEL 2004] Zeller, M., et al., Predictors of attrition from a pediatric weight management program. J Pediatr, 2004. pp. 466-470.

[GOM 2009] Gomez V, Abasolo JE. Using data mining to describe long hospital stays. Paradigma 2009. pp. 1-10.

[SON 2010] Son YJ, Kim HG, Kim EH, Choi S, Lee SK. Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients. Healthc Inform Res 2010. pp. 253-259.

[CHO 2010] Choi K, Chung S, Rhee H, Suh Y. Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers. Healthc Inform Res 2010. pp. 67-76.

[PAN 2013] Pan YJ, Liu SK, Yeh LL. Factors affecting early attrition and later treatment course of antidepressant treatment of depression in naturalistic settings: An 18-month nationwide population-based study. J Psychiatr Res 2013. pp. 916-925.

[USH 2011] Usha Rani, K. Analysis of heart diseases dataset using neural network approach. International Journal of Data Mining & Knowledge Management Process (IJDKP) 2011. pp. 1-8.

[KAN 2009] Kang JO, Chung SH, Suh YM. Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques. J Kor Soc Med Informatics 2009. pp. 13-23.

[LIF 2011] Li F, Lei J, Tian Y, Punyapatthanakul S, Wang YJ. Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking. Australian Computer Society, 2011. pp. 119-124.

[JDM 2014] The Data Mining Java Api-Oracle
http://docs.oracle.com/cd/B28359_01/datamine.111/b28131/java_api.htm#BGBBBDEA
Último acceso: abril 2014.

[RAP 2014] RapidMiner
http://rapidminer.com/

Último acceso: abril 2014.

[SAS 2014] SAS
http://www.sas.com/en_us/home.html
Último acceso: abril 2014.

[WEK 2014] Weka
http://www.cs.waikato.ac.nz/ml/weka/
Último acceso: abril 2014.

[RHA 2014] R and Data Mining
http://www.rdatamining.com/tutorials/rhadoop
Último acceso: abril 2014.

[ERW 2014] CA ERwin
http://www.erwin.com/
Último acceso: agosto 2014.

[PEN 2014] Pentaho
http://www.pentaho.com/
Último acceso: agosto 2014.

[IBM 2005] IBM. 2005. DataMart Consolidation: Getting Control of your Entreprise Information, 2005. Julio.

[KIM 2002] Kimball, R AND Ross, M. The DataWarehouse Toolkit. 2002: John Wiley & Sons, Inc .

[HAN 2006] Han, J., M. Kamber, and J. Pei, Data mining: concepts and techniques. 2006: Morgan kaufmann.

[MIN 2013] Ministerio de Salud del Perú. 2013. Análisis de la Situación del cáncer en el Perú, 2013. Noviembre

[IAN 2011] Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. 2011: Morgan Kauffman Publishers Inc.