UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

FACULTAD DE INGENIERÍA

ESCUELA DE FORMACION PROFESIONAL DE INGENIERIA DE SISTEMAS Y COMPUTACIÓN

TESIS

Reconocimiento de patrones en enfermedades respiratorias mediante minería de datos para mejorar el diagnostico en pacientes del Hospital Daniel Alcides Carrión – Pasco

Para optar el título profesional de:

Ingeniero de Sistemas y Computación

Autor: Bach. Henry Herzen CURI ESTRELLA

Asesor: Mg. Percy RAMIREZ MEDRANO

Cerro de Pasco- Perú - 2020

UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

FACULTAD DE INGENIERÍA

ESCUELA DE FORMACIÓN PROFESIONAL DE INGENIERIA DE SISTEMAS Y COMPUTACIÓN

TESIS

Reconocimiento de patrones en enfermedades respiratorias mediante minería de datos para mejorar el diagnostico en pacientes del Hospital Daniel Alcides Carrión – Pasco

Sustentada y aprobada ante los miembros del jurado:

_____	_____
Dr. Ángel Claudio NUÑEZ MEZA	Mg. Raúl Delfín CONDOR BEDOYA
PRESIDENTE	MIEMBRO

_____

Mg. Oscar Clevorio CAMPOS SALVATIERRA

MIEMBRO

DEDICATORIA

Dedico esta tesis A Dios quien inspiro mi espíritu para la conclusión de esta tesis. A mis padres quienes me dieron educación, apoyo y consejos. A mis maestros, quienes sin su ayuda no hubiera podido hacer esta tesis.

RECONOCIMIENTO

Agradezco a nuestra alma mater la Universidad Nacional Daniel Alcides Carrión por habernos acogido en sus aulas, a nuestros queridos docentes y a mi asesor quien me apoyó en el desarrollo del trabajo de investigación.

RESUMEN

En cumplimiento a las disposiciones vigentes del Reglamento de Grados y Títulos de nuestra Facultad de Ingeniería, Escuela de Formación Profesional de Sistemas y Computación, pongo a vuestra consideración la presente Tesis Intitulado “RECONOCIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS MEDIANTE MINERÍA DE DATOS PARA MEJORAR EL DIAGNOSTICO EN PACIENTES DEL HOSPITAL DANIEL ALCIDES CARRIÓN - PASCO”, con el propósito de optar el título profesional de Ingeniero de Sistemas y Computación.

En la actualidad en el hospital Daniel Alcides Carrión de Pasco las enfermedades respiratorias han ido mutando considerablemente, siendo esta una causa para que los galenos confundan su diagnóstico, o que no se tenga el informe de manera oportuna para la toma de decisiones, es por eso que esta investigación busca mediante un modelo de reconocimiento de enfermedades respiratorias diagnosticar si el paciente presenta o no cuadro de neumonía, ya que esta es la infección respiratoria aguda que en mayor porcentaje afecta a la región Pasco.

No dudo pues, que esta tesis sea un aporte significativo que contribuya al desarrollo académico universitario, así como al de las empresas de nuestra región.

Palabras clave: Minería de datos, reconocimiento de patrones

EL AUTOR.

ABSTRACT

In compliance with the current regulations of the Degree and Degrees Regulation of our Faculty of Engineering, School of Systems and Computing Professional Training, I put to your consideration the present Thesis “RECONOCIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS MEDIANTE MINERÍA DE DATOS PARA MEJORAR EL DIAGNOSTICO EN PACIENTES DEL HOSPITAL DANIEL ALCIDES CARRIÓN - PASCO” , with the purpose of choosing the professional title of Systems and Computing Engineer.

Currently, at the Daniel Alcides Carrión de Pasco hospital, respiratory diseases have been mutating considerably, this being a cause for doctors to confuse their diagnosis, or not having the report in a timely manner for decision-making, that is why This research seeks to diagnose whether the patient has pneumonia or not, because this is the acute respiratory infection that most affects the Pasco region.

I do not doubt, then, that this thesis is a significant contribution that contributes to university academic development, as well as that of the companies in our region.

Keywords: Data mining, pattern recognition

THE AUTHOR

INTRODUCCIÓN

Hoy en día el uso de los sistemas de información para la toma de decisiones mediante minería de datos apoya en la administración, procesamiento y distribución de la información en una organización, hasta el punto que se ha hecho cada vez más indispensable. Estos sistemas permiten lograr ahorros significativos en tiempo y mano de obra y costos, debido a que automatizan tareas operativas de la organización y ofrecen un gran apoyo en el proceso de toma de decisiones, esto permite a su vez lograr ventajas competitivas al momento de la implementación y uso del sistema de información.

En este proyecto de tesis de grado se pretende modelar un reconocedor de patrones en enfermedades respiratorias mediante minería de datos que mejore el diagnostico en pacientes del hospital Daniel Alcides Carrión - Pasco. La información recolectada para este fin procede de historias clínicas de pacientes en el año 2016, información sensible y confidencial por lo que se tuvo que establecer un criterio ético de no divulgación y/o uso indebido de la data consignada en estos documentos. Esta situación fue un limitante crítico para que se haya desarrollado esta tesis.

La investigación, consta de 5 capítulos, a groso modo se describe cada uno de ellos:

En el Capítulo I, se establece el problema de la investigación, formulando los problemas, objetivos, la justificación del estudio entre otros de la investigación y otros que amerite de acuerdo al esquema vigente de tesis.

En el Capítulo II, se presenta el marco teórico, comprende los antecedentes y las

bases teóricas utilizadas para el desarrollo del estudio, poniendo principal atención en las variables planteadas: reconocimiento de patrones de enfermedades respiratorias y diagnóstico de pacientes en el hospital Daniel Alcides Carrión de Pasco, así como la definición de términos necesarios para un mejor entendimiento de la tesis, la formulación de hipótesis y finalmente las variables de investigación y sus indicadores.

En el Capítulo III, se describe la metodología y técnicas de investigación propuesta y utilizada para el proceso de la investigación.

En el Capítulo IV, abarca el análisis de resultados y discusión de la investigación, en función del modelo y los resultados obtenidos de la minería de datos.

Creo que la presente investigación será un aporte significativo tanto a la institución educativa, la universidad y la sociedad.

EL AUTOR

INDICE

ÍNDICE DE FIGURAS

CAPITULO I

PROBLEMA DE INVESTIGACIÓN

1.1. Identificación y determinación del problema

Los errores médicos son objetos de estudio a nivel internacional desde algunas décadas atrás, en el año 1999 el Instituto de Medicina de los Estados Unidos público un informe titulado “To Err is human: Building a Safer Health System” en 1984, tocando temas que fueron escasamente discutidos en ese tiempo: la seguridad del paciente, donde detallaba que 98000 personas murieron por errores médicos evitables, siendo la sexta causa de muerte en el mencionado país. Este problema aqueja a nivel global y lo que se busca desde hace muchos años es tener una cultura que permita cambiar los estilos de trabajos, promover políticas y estrategias para reducir el impacto de los errores.

Esta situación también se presenta en el Hospital Daniel Alcides Carrión de

Pasco donde se manejan un sinfín de información, al terminar el día tienen decenas de facturas, prestamos, atenciones, servicios que son registrados y almacenados. Todos esos datos pueden contribuir a brindar conocimiento a la organización por ejemplo se puede conocer cuál es la enfermedad más frecuente por zonas o por edades. Para ello necesitan de una herramienta que le permita manejar todos los datos y convertirlo en conocimiento, utilizando patrones, búsquedas heurísticas entre otros. Con el procesamiento de datos, como el de las historias clínicas, mediante minería de datos, se podría obtener patrones sobre síntomas y enfermedades. La situación del sector salud en el Perú con respecto al uso de minería de datos como herramienta predictiva no son muchas, si bien existen iniciativas tecnológicas, como reglamentar las historias clínicas de forma electrónicas, o crear sistemas de información que ayuden a los procesos de las entidades de salud, falta cosas por hacer. Así pues, con esta investigación se busca demostrar la gran ayuda que pueden brindar los modelos que muestren un patrón de comportamiento de los pacientes y sirva como base para reorientar los recursos.

1.2. Delimitación de la investigación

La delimitación de la investigación se presenta desde las siguientes perspectivas:

- **Relevancia social:** El presente trabajo de investigación beneficiara a la comunidad de Yanacancha y alrededores.

- **Implicaciones prácticas:** La investigación permite poner en práctica los conocimientos adquiridos durante la formación profesional dentro de la materia inteligencia artificial y base de datos, permitiendo solucionar los problemas existentes en una organización como el caso planteado.

- **Utilidad metodológica:** Mediante la investigación se genera un sistema de información transaccional que puede a su vez ser utilizado por otras instituciones que requieran automatizar y hacer más eficiente sus procesos internos de control de asistencia.

1.3. Formulación del problema

1.3.1. Problema general

¿Cómo influye el reconocimiento de patrones en enfermedades respiratorias mediante minería de datos en el diagnóstico de pacientes del Hospital Daniel Alcides Carrión - Pasco?

1.3.2. Problemas específicos

1 ¿Qué variables de enfermedades respiratorias se relacionan con el diagnóstico de pacientes del Hospital Daniel Alcides Carrión - Pasco?

2 ¿En qué medida la aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco?

1.4. Formulación de objetivos

1.4.1. Objetivo general

Determinar la influencia de reconocer patrones en enfermedades respiratorias mediante minería de datos en el diagnóstico de pacientes del hospital Daniel Alcides Carrión – Pasco.

1.4.2. Objetivos específicos

1 Identificar las variables de enfermedades respiratorias que se relacionan con el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

2 Establecer el modelo sobre enfermedades respiratorias mediante minería de datos para predecir el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco.

1.5. Justificación de la investigación

Las organizaciones de la actualidad son juzgadas no únicamente por la calidad de sus productos o servicios, sino también por el grado en el que comparten información con sus clientes, empleados y socios. Sin embargo, la gran mayoría de las organizaciones tienen una abundancia de datos, pero sin ser aprovechados no conocidos. Es por ello que se desarrollará la minería de datos, la misma que trata de obtener de ellos no solo información, sino una verdadera inteligencia que le confiera a la organización una ventaja competitiva, para la toma de decisiones

En la actualidad en el hospital Daniel Alcides Carrion de Pasco las enfermedades respiratorias han ido mutando considerablemente, siendo esta una causa para que los galenos confundan su diagnóstico, o que no se tenga el informe de manera oportuna para la toma de decisiones.

Es por eso que con esta investigación y el modelo resultante se busca apoyar a los médicos del hospital en el oportuno y debido diagnóstico para una adecuada toma de decisiones en cuanto a la atención médica.

1.6. Limitaciones de la investigación

Dadas las características del tema y del trabajo de investigación propuesto, se establecieron tres tipos de limitación:

Limitación temporal: el estudio se realizó sobre datos recogidos durante el periodo de junio a agosto del 2019.

Limitación de información: Las limitaciones encontradas para la realización de esta investigación se centran en el recelo de los responsables en compartir los datos y base de datos de los pacientes con enfermedades respiratorias, ya que como comprenderán es información confidencial, por lo que se optó por mantener anónima la información de cada paciente. De igual modo existe una limitación económica, ya que todo el proyecto investigativo es autofinanciado.

Limitaciones económicas: El estudio realizado es autofinanciado, existiendo por consiguiente limitaciones en el gasto de los recursos necesarios para lograr los objetivos de la investigación.

CAPITULO II

MARCO TEÓRICO

2.1 Antecedentes de estudio

El interés en la presente investigación es en parte despertada por las investigaciones realizadas en este particular, analizando lo que se ha venido haciendo dentro del dominio de la minería de datos y el sector salud, a continuación, se mencionan antecedentes de tipo internacional, nacional y local de ser el caso.

Internacional

Eapen, A. con la investigación “Application of data mining in medical applications”, tesis de maestría, realizado el 2004, Universidad de Waterloo, Canada. El estudio indica que una cantidad tan grande de datos no puede ser procesada por humanos en poco tiempo para hacer

diagnósticos, pronósticos y programas de tratamiento. Un objetivo principal de esta tesis es evaluar las herramientas de minería de datos en aplicaciones médicas y de atención médica para desarrollar una herramienta que pueda ayudar a tomar decisiones oportunas y precisas. Se consideran dos bases de datos médicas, una para describir las diversas herramientas y la otra como estudio de caso. La primera base de datos está relacionada con el cáncer de mama y la segunda está relacionada con el conjunto mínimo de datos para la salud mental (MDS-MH). Los resultados indicaron que, para el estudio de caso principal, es decir, el problema de salud mental, es posible obtener resultados precisos de más del 70 al 80% de precisión.

Así mismo está el estudio de Vela, Y. “Caracterización epidemiológica de las infecciones respiratorias agudas (IRA) en hospitalización pediátrica, clínica Antioquia-Bello, Colombia, año 2016”, tesis de maestría, año 2018, Universidad Nacional Autónoma de Nicaragua. La tesis tiene como objetivo caracterizar de forma epidemiológica las infecciones respiratorias agudas denominadas IRA, describiendo las características sociodemográficas, clínico epidemiológicas y conociendo el comportamiento epidemiológico de los casos de IRA según los días de hospitalización. Las principales conclusiones y resultados fueron que la población más vulnerable a este tipo de infecciones es la lactante menor (109 casos), seguida de la preescolar (108 casos). El 97% de la población procede del área urbana, y solo un 3% del área rural. Se revisa en el estudio que haber padecido antes IRA no podría ser un factor que incida en su padecimiento. La IRA más común en la Clínica Antioquia en

su sede Bello es la Neumonía con un 60%, seguida por la bronquiolitis con 21%, por otro lado, solo un 7% de las hospitalizaciones fueron por afecciones respiratoria del tracto bajo. Los agentes causantes de las IRA registrados en las historias clínicas, fueron atípicos (127 casos), micoplasma (74 casos), Virus Sincitial Respiratorio (34 casos).

Nacional

A nivel nacional tenemos la investigación realizada por Bernuy, A., "Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la escuela profesional de ingeniería de computación y sistemas, universidad de San Martín de Porres, Lima-Perú", tesis de pregrado, en el año 2018, Universidad San Martin de Porras. Lima. Este estudio se realizó con datos de 1304 ingresantes que fueron clasificados en tres factores: sociales, económicos y académicos. Se realizaron predicciones a través de tres técnicas: regresión lineal, árbol de decisiones y support vector machines, y el mejor resultado de 82.87% se obtuvo utilizando árbol de decisiones. De los diferentes factores, los que más influyeron en el rendimiento académico fueron los siguientes: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios. Utilizando minería de datos fue posible realizar predicciones del rendimiento académico de los ingresantes. Esto permitió la detección de ingresantes que podrían enfrentarse a problemas en sus estudios.

En el trabajo de investigación de Candela, C., "Proceso de Descubrimiento de Conocimiento para Predecir el Abandono de

Tratamiento en una Entidad de Salud Pública" realizado en el 2015, en la ciudad de Lima, cuyo objetivo fue el de 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. Realizo pruebas con cuatro algoritmos dando como resultado: "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%."Concluyendo que, gracias al algoritmo SVM, se pudo determinar los factores más influyentes como son la edad, la autoestima, los medicamentos suministrados, entre otros y, gracias al algoritmo de árbol de decisión, las reglas asociadas a las categorías de tiempo de duración de la hospitalización. Esta investigación contribuye al uso de minería de datos, el beneficio de emplearlo y permite conocer el funcionamiento de los algoritmos para a partir de ello realizar el modelo propuesto.

Se encuentra la investigación realizada por Daza, A. "Un modelo basado en arboles de decisión para predecir la deserción estudiantil en la educación superior privada", año 2016, Universidad Cesar Vallejo. En ella indica que las técnicas de minería de datos permiten obtener información útil que se encuentra oculta en grandes bases de datos que en su mayoría solo son usados para realizar operaciones transaccionales, así como archivos que aún no han sido ingresado a las bases de datos. Debido a la gran cantidad de datos que tienen las Instituciones de Educación Superior Universitaria Cesar Vallejo propone hacer uso de las

técnicas de minería de datos para predecir la deserción o el abandono en la Educación Superior Privada. Para el desarrollo de proyecto se usó la metodología CRIPS-DM con la herramienta comercial Spss Clementine 12.0, para los cuales se hicieron uso de la técnica de minería de datos árboles de decisión sobre 1761 datos de los estudiantes de la Escuela profesional de Ingeniería de Sistemas con 27 atributos para cada uno de ellos. La conclusión a la que se llego es que aplicando minería de datos con técnicas de árboles de decisión se podía predecir el comportamiento de la deserción estudiantil en un 89%.

A nivel local no se encontraron investigaciones relacionadas con esta temática.

2.2 Bases teóricas - científicas

2.2.1 Enfermedades respiratorias

El tracto respiratorio superior incluye la nariz, la boca, las fosas nasales, la faringe y la laringe, y en el tracto respiratorio inferior se encuentra la tráquea, bronquio principal y pulmón (Figura 2.1).

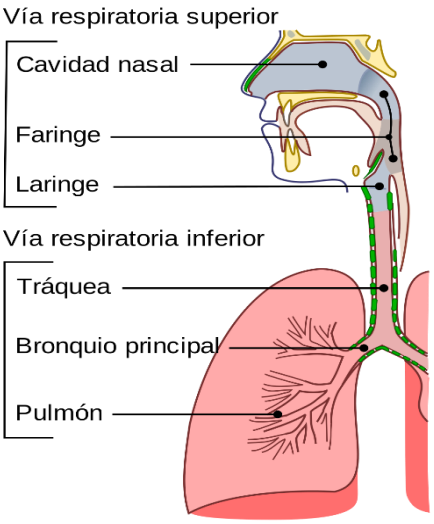


Figura 2.1. Componentes de las vías respiratorias

Estas estructuras dirigen la respiración del aire desde el exterior hacia los pulmones para que tenga lugar la respiración. Una infección aguda del tracto respiratorio, a la que denominaremos para la investigación cono “enfermedad respiratoria”, es un proceso infeccioso de cualquiera de los componentes de la vía aérea superior o inferior (Lambert et al., 2008). La infección de las áreas específicas del tracto respiratorio superior se puede nombrar específicamente. Ejemplos de estos pueden incluir rinitis (inflamación de la cavidad nasal), infección sinusal (sinusitis o sinusitis de rinoceronte) - inflamación de los senos ubicados alrededor de la nariz, resfriado común (nasofaringitis) - inflamación de la faringe, hipofaringe, úvula y amígdalas, faringitis (inflamación de la faringe, la úvula y las amígdalas), epiglotitis (inflamación de la porción superior de la laringe o la epiglotis), laringitis (inflamación de la laringe), laringotraqueitis (inflamación de la laringe y la tráquea), y traqueitis (inflamación de la

tráquea). Las infecciones de las vías respiratorias superiores son una de las causas más frecuentes de visitas médicas con síntomas variables que van desde secreción nasal, dolor de garganta, tos, dificultad para respirar y letargo (Williams et al., 2002; Lambert et al., 2008).

Clasificación de las enfermedades respiratorias

Las enfermedades respiratorias o infecciones respiratorias agudas (IRA) se clasifican como infecciones del tracto respiratorio superior (ITRS) o infecciones del tracto respiratorio inferior (ITRI). El tracto respiratorio superior consiste en las vías respiratorias desde las fosas nasales hasta las cuerdas vocales en la laringe, incluidos los senos paranasales y el oído medio. El tracto respiratorio inferior cubre la continuación de las vías respiratorias desde la tráquea y los bronquios hasta los bronquiolos y los alvéolos (Simoes et al., 2006).

Causas de las infecciones respiratorias agudas

Las infecciones respiratorias agudas son causadas principalmente por virus, pero se han confirmado las etiologías bacterianas. Un estudio para la Organización Mundial de la Salud de casos controlados de pacientes de padecimiento general de IRA descubrió que los virus representaban el 58% de las infecciones agudas del tracto respiratorio y las bacterias como el estreptococo del Grupo A

era responsable del 11%, y el 3% de los pacientes que tenían infecciones bacterianas y virales mixtas.

Transmisión de las infecciones respiratorias agudas

La transmisión se realiza a través de gotitas respiratorias o por manos contaminadas con virus. La inflamación de la mucosa del tracto respiratorio superior (nariz, garganta, senos paranasales) aumenta las secreciones, provoca estornudos y tos que facilitan la propagación. En muchos países en desarrollo y desarrollados, las IRA son la enfermedad infecciosa más común en la población general. Las infecciones agudas del tracto respiratorio son las principales causas de morbilidad y mortalidad en los países desarrollados y en desarrollo. La importancia de las infecciones agudas del tracto respiratorio en los adultos que trabajan y en los ancianos ha sido reconocida por mucho tiempo, y la mayor parte de los esfuerzos de investigación y programas de prevención se han dirigido a estos grupos. Más recientemente, se ha renovado la atención dada al impacto que las infecciones del tracto respiratorio tienen en bebés y niños y las perspectivas de prevención en este grupo.

2.2.2 Reconocimiento de patrones en enfermedades respiratorias

Las enfermedades respiratorias generalmente presentan un patrón (estándar o modelo) de comportamiento mediante el cual a priori un médico puede diagnosticar si un paciente sufre de este cuadro, ya

sea como infección del tracto respiratorio inferior o del tracto respiratorio superior. Eccles et al. (2007) indica los siguientes:

Infecciones del tracto respiratorio inferior (ITRS o IRS).

El tracto respiratorio inferior es la parte del tracto respiratorio debajo de las cuerdas vocales. Aunque a menudo se usa como sinónimo de neumonía, la rúbrica de la infección del tracto respiratorio inferior también se puede aplicar a otros tipos de infección, como el absceso pulmonar y la bronquitis aguda. Los síntomas incluyen:

- Dificultad para respirar
- Debilidad
- Fiebre alta
- Tos y fatiga

Las infecciones del tracto respiratorio inferior ejercen una presión considerable en el presupuesto de salud y generalmente son más graves que las infecciones del tracto respiratorio superior. Desde 1993 ha habido una ligera reducción en el número total de muertes por infección del tracto respiratorio inferior, representaron 3.9 millones de muertes en todo el mundo según la Organización Mundial de la Salud. Hay una serie de infecciones agudas y crónicas que

pueden afectar el tracto respiratorio inferior. Las dos infecciones más comunes son bronquitis y neumonía.

Infecciones del tracto respiratorio superior (ITRS o IRS).

Son las enfermedades causadas por una infección aguda que afecta el tracto respiratorio superior: nariz, senos nasales, faringe o laringe. Esto comúnmente incluye: amigdalitis, faringitis, laringitis, sinusitis, otitis media y resfriado común. Las infecciones agudas del tracto respiratorio superior incluyen rinitis, faringitis / amigdalitis y laringitis, a menudo denominadas resfriado común, y sus complicaciones: sinusitis, infección del oído y a veces bronquitis (aunque los bronquios generalmente se clasifican como parte del tracto respiratorio inferior). Los síntomas de IRS comúnmente incluyen:

- Tos
- Dolor de garganta
- Secreción nasal
- Congestión nasal
- Dolor de cabeza
- Fiebre baja
- Presión facial y

- Estornudos.

El inicio de los síntomas generalmente comienza de 1 a 3 días después de la exposición a un patógeno microbiano. La enfermedad generalmente dura de 7 a 10 días. La faringitis estreptocócica hemolítica beta o amigdalitis del grupo A generalmente se presenta con un inicio repentino de dolor de garganta, dolor al tragar y fiebre. La amigdalitis no suele causar secreción nasal, cambios en la voz o tos.

2.2.3 Minería de datos

El desarrollo de la tecnología de la información ha generado una gran cantidad de bases de datos y enormes datos en diversas áreas. La investigación en bases de datos y tecnología de la información ha dado lugar a un enfoque para almacenar y manipular estos datos valiosos para una mayor y mejor toma de decisiones. La minería de datos es un proceso de extracción de información útil grandes datos. También se le llama proceso de descubrimiento de conocimiento, extracción de conocimiento a partir de datos, extracción de conocimiento o análisis de datos/patrones (figura 2.2).

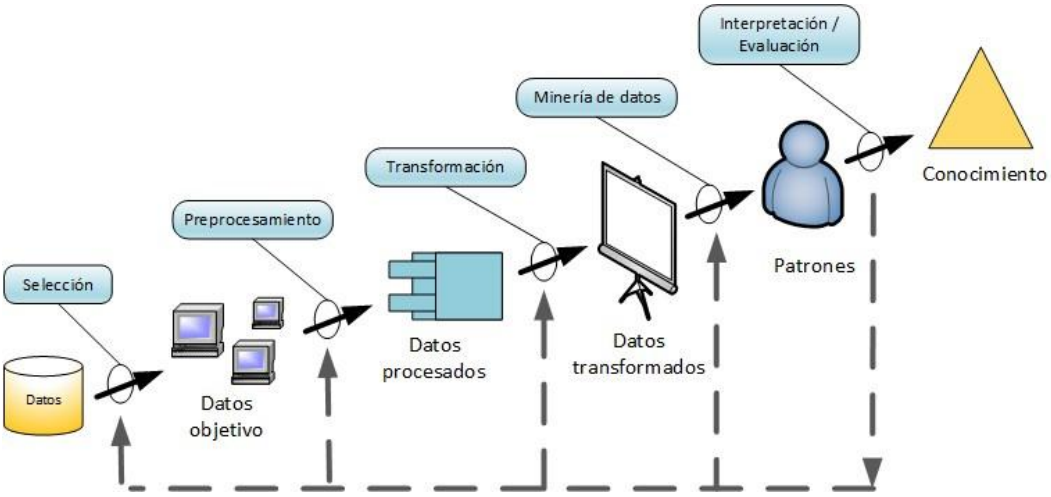


Figura 2.2. Proceso de descubrimiento de conocimiento

Una definición de minería de datos la obtenemos de Han, J. et. al. (2012).

> “La minería de datos es el proceso de descubrir patrones interesantes a partir de grandes cantidades de datos. Como proceso de descubrimiento de conocimiento, generalmente implica la limpieza de datos, la integración de datos, la selección de datos, la transformación de datos, el descubrimiento de patrones, la evaluación de patrones y la presentación de conocimientos”.

Shahiri, A. et al. (2015), define la minería de datos como el proceso de analizar datos ya existentes en una base de datos para encontrar patrones ocultos en ella. Los patrones extraídos deben ser significativos en cierto sentido, por ejemplo, económicamente o desde el punto de vista de la seguridad del paciente.

Además del término minería de datos, existen otros términos más o menos equivalentes utilizados en la literatura, como el aprendizaje automático, el análisis predictivo y el descubrimiento de conocimiento en bases de datos (KDD). La minería de datos convierte una gran colección de datos en conocimiento. Un motor de búsqueda (por ejemplo, Google) recibe cientos de millones de consultas todos los días. Cada consulta se puede ver como una transacción en la que el usuario describe su necesidad de información. ¿Qué conocimiento novedoso y útil puede aprender un motor de búsqueda de una colección tan enorme de consultas recopiladas de los usuarios a lo largo del tiempo? Curiosamente, algunos patrones encontrados en las consultas de búsqueda de usuarios pueden revelar un conocimiento invaluable que no puede obtenerse leyendo solo elementos de datos individuales. Por ejemplo, las Tendencias de la gripe de Google utilizan términos de búsqueda específicos como indicadores de la actividad de la gripe. Encontró una estrecha relación entre la cantidad de personas que buscan información relacionada con la gripe y la cantidad de personas que realmente tienen síntomas de gripe. Este ejemplo muestra cómo la minería de datos puede convertir una gran colección de datos en conocimiento que puede ayudar a enfrentar un desafío global actual.

Ciclo de la minería de datos

Según MacLennan, J. et al (2008) el ciclo del proyecto de minería de datos contiene aproximadamente los siguientes

pasos, aunque varía según el propósito de la minería de datos

1. Formación de problemas comerciales. ¿Cuáles son los problemas a resolver? ¿Cómo pueden resolverse estos problemas con la minería de datos? ¿Qué tarea de minería de datos es adecuada?

2. Preprocesamiento de datos. El paso de preprocesamiento de datos contiene, por ejemplo, la recopilación de todos los datos relevantes en un solo lugar, la limpieza de datos para eliminar el ruido y los datos redundantes, la agrupación y la agregación de atributos discretos. El propósito es transformar los datos sin procesar para que sean útiles para la minería. Este es el paso que consume más recursos.

3. Modelo de construcción. Se eligen uno o varios algoritmos de minería de datos dependiendo de la tarea de minería de datos y se construyen los modelos. Por lo general, se construyen varios modelos con diferentes algoritmos o con diferentes parámetros de algoritmos para poder comparar su desempeño en la descripción de datos o predicciones.

4. Evaluación del modelo. Se evalúa el rendimiento predictivo de los modelos y se inspeccionan y evalúan los patrones revelados en términos de utilidad y valor para el área de estudio.

5. Predicción. En muchos casos, la predicción es el objetivo de la minería de datos. Los modelos que han sido entrenados con datos en el paso de construcción del modelo ahora pueden usarse para predecir nuevos casos de datos.

6. Integración de aplicaciones. Incrustar la minería de datos en la aplicación empresarial.

Estructura y modelo

Dos conceptos importantes en la minería de datos son la estructura de minería de datos y el modelo de minería de datos. Una estructura de minería de datos define la forma del problema de minería de datos. Contiene información sobre las columnas de datos incluidas, como el género y la edad, incluidos sus tipos de datos y si son discretos (es decir, tienen un número establecido de valores, como el sexo) o continuos (es decir, son numéricos, como la edad). La estructura de minería contiene los datos de origen que se utilizan para la capacitación y las pruebas de los modelos de minería e información sobre la cantidad de datos que se deben utilizar para la capacitación y las pruebas. El modelo de minería de datos transforma las filas de datos de origen en casos y realiza minería de datos con estos. Utiliza algunos o todos los datos de origen, según los filtros aplicados al modelo. El modelo de minería de datos utiliza un algoritmo de minería de datos y algunas o todas las columnas de la estructura de

minería como atributos de minería de datos y especifica si estos atributos se utilizarán como entrada, salida o ambas. El modelo luego usa las entradas para aprender sobre las salidas. La idea detrás de la minería de datos es mostrar ejemplos de datos de un modelo de minería de datos, que contengan tanto entradas como salidas, de las cuales puede extraer patrones. Esto se denomina fase de capacitación. Luego, los patrones pueden estudiarse por sí mismos o aplicarse a nuevos ejemplos de datos. Para probar el rendimiento predictivo, el modelo entrenado solo recibe los atributos de entrada de los nuevos casos y, a partir de ellos, intenta predecir el estado del atributo de salida. La predicción se compara con el estado conocido del atributo de salida de ese caso. De esta manera, se puede evaluar el rendimiento del modelo en la predicción de los resultados. Esto se llama la fase de prueba.

Tipos de algoritmos en minería de datos

Hay varios algoritmos de minería de datos que se pueden aplicar para resolver un problema. Dependiendo de la naturaleza del problema, se pueden combinar una o más tareas diferentes para resolverlo porque todos funcionan de manera diferente. Las tareas generales de minería de datos incluyen clasificación, agrupamiento, asociación, regresión, proyección previa, análisis de secuencia y análisis de

desviación. La elección del algoritmo de minería de datos luego decide exactamente cómo se analizan los datos para los patrones y en qué forma son los patrones identificados (por ejemplo, árboles y reglas). A continuación, se presentan las tipos de algoritmos que se pueden aplicar en este estudio, es decir, la clasificación y el análisis de asociación. descripciones de los algoritmos que se pueden usar para realizar cada tarea.

Clasificación

La clasificación es la tarea de asignar un estado al atributo de salida de cada caso. La tarea de clasificación consiste en describir los patrones del atributo de salida en términos de los atributos de entrada. Un modelo se entrena con datos de entrenamiento donde se conoce el atributo de salida, y luego se puede usar para clasificar el atributo de salida en los datos de prueba. La clasificación se llama una tarea supervisada porque requiere un objetivo supuesto, el atributo de salida, para obtener el resultado. Los algoritmos de clasificación disponibles son (MacLennan, 2008): Naive Bayes, árboles de decisión, redes neuronales y regresión logística, que se describen a continuación.

- **Algoritmo de Bayes.** Naive Bayes es el más simple de los algoritmos de clasificación. Construye patrones contando las correlaciones entre todos los diferentes

estados de los atributos de entrada y todos los diferentes estados de los atributos externos. Los atributos solo pueden tener valores discretos. Naive Bayes se basa en los teoremas de Bayes y es ingenuo en el sentido de que no tiene en cuenta las posibles dependencias entre los atributos de entrada. Las dependencias fuertes entre el atributo de entrada pueden, por lo tanto, sesgar los patrones identificados. Naive Bayes a menudo se usa al comienzo del proceso de minería de datos para explorar rápidamente los datos, pero también puede ser un poderoso predictor en algunas situaciones. Algoritmos más avanzados como árboles de decisión y redes neuronales se usan típicamente para la predicción cuando están disponibles. Los patrones generados por Naive Bayes incluyen los llamados atributos característicos que pueden interpretarse como los principales influenciadores. Las características del atributo se expresan como una combinación de atributo-estado con una frecuencia asociada que indica la proporción de casos con el estado de salida de destino que también tenía esta combinación específica de atributo de entrada de estado. La frecuencia puede interpretarse como la fuerza de la influencia en el resultado de la infección.

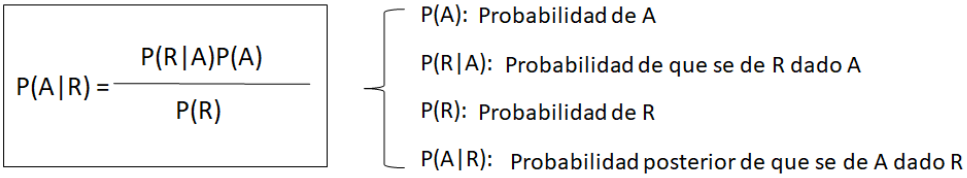


Figura 2.3. Algoritmo Naive Bayes basado en las probabilidades

- **Algoritmo de árboles de decisión.** Los árboles de decisión pueden manejar atributos discretos y continuos, pero agrupa los valores continuos si corresponde. El algoritmo funciona de manera recursiva para construir un árbol que luego puede usarse para la predicción. Busca el atributo de entrada que divide más claramente los datos entre los estados del atributo de salida. Ese atributo de entrada se utiliza para dividir los datos en subconjuntos y luego se repite el mismo procedimiento para cada subconjunto y así sucesivamente. Cuando se clasifica un nuevo caso de datos, se compara con las divisiones del árbol construido, creando así una ruta desde la raíz hasta un nodo hoja. Ese nodo hoja contiene el estado predicho del atributo de salida. Durante el entrenamiento, el árbol se poda utilizando dos parámetros de algoritmo para que el árbol resultante no sea demasiado profundo, lo que puede causar un entrenamiento excesivo. Los árboles muy profundos tienden a sobrerrepresentar los datos de entrenamiento en lugar de generalizar las reglas, lo que puede resultar en un mal desempeño al clasificar nuevos

casos de datos. El arbol de decisión es uno de los algoritmos más populares porque es rápido, fácil de entender y preciso si se usa correctamente.

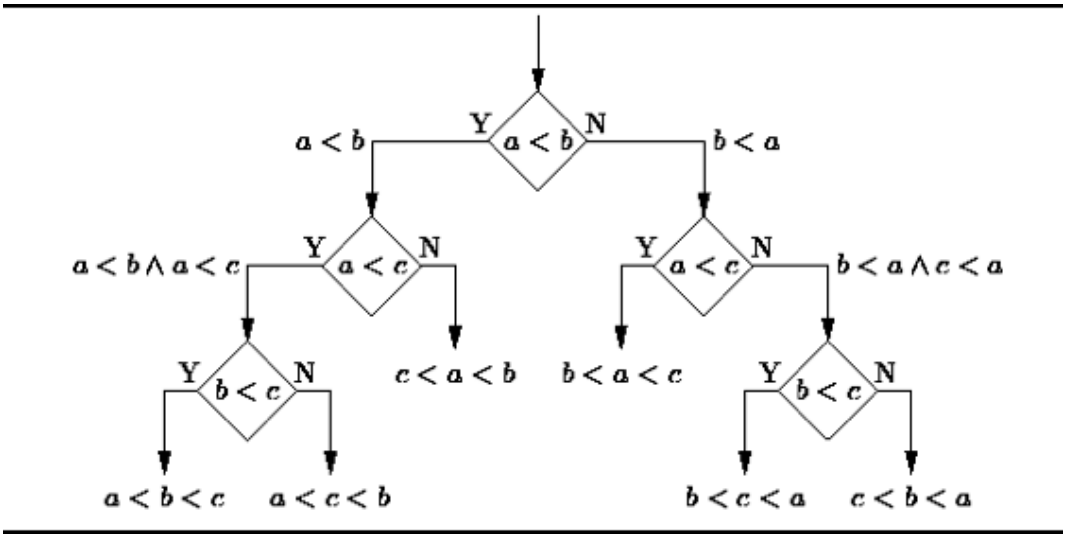
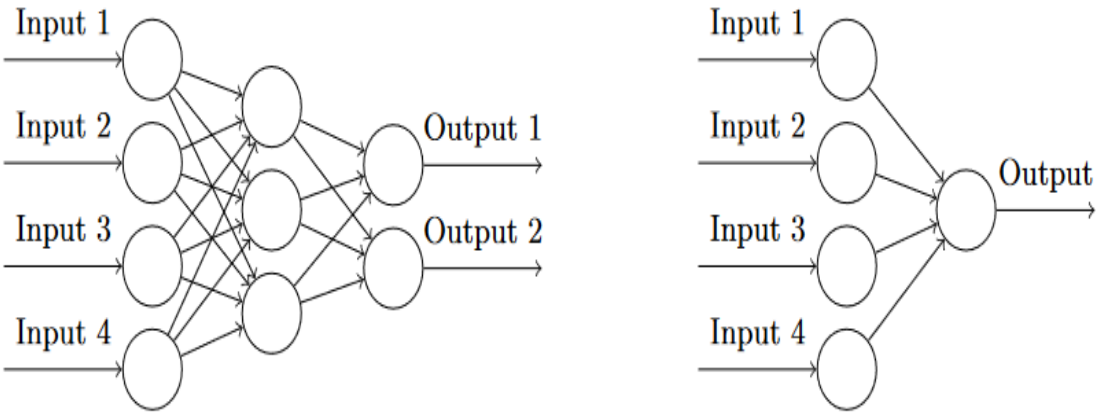


Figura 2.3. Ejemplo de Árbol de decisión.

- **Algoritmo de red neuronal.** El algoritmo de red neuronal es una red neuronal artificial que imita la forma en que funciona la mente humana cuando se le presenta un problema. Analiza todas las combinaciones posibles de entradas y salidas y asigna pesos a sus relaciones. También busca combinaciones de entradas que se correlacionan con una salida, aunque las entradas por sí solas no. También hay una capa oculta de nodos entre las entradas y las salidas, de modo que las entradas no tienen que estar directamente correlacionadas con una salida. Figura 2.4a.

- **Algoritmo de regresión logística.** La regresión logística es un caso especial del algoritmo de la red neuronal en la forma en que no contiene capa oculta, vea la Figura 2.4b, pero además de eso son idénticos y, por lo tanto, se comportan de manera similar. La capa oculta eliminada no necesariamente lo convierte en un algoritmo de debilitamiento a la hora de predecir nuevos casos de datos. En algunas situaciones, incluso puede funcionar mejor que la Red Neural porque la complejidad reducida implica menos riesgo de sobre entrenamiento. Tanto la red Neural como la regresión logística pueden manejar atributos discretos y continuos



a) Red neural b) regresión logística

Figura 2.4. Representación de red neuronal y de regresión logística.

Análisis de asociación

La asociación implica analizar conjuntos de elementos para encontrar conjuntos de elementos que ocurren con frecuencia

y, a partir de estos, formular reglas de asociación. La minería de datos de la asociación a menudo se denomina "análisis de cesta de compra" porque aumentar las ventas cruzadas al analizar los comportamientos de compra de los clientes en forma de tablas de transacciones es una aplicación común. El análisis de asociación es principalmente una tarea descriptiva, es decir, su propósito es describir patrones en los datos, pero también es posible usarlo para predecir el atributo de salida.

Las Reglas de asociación simplemente cuentan los estados de atributo de entrada y salida y la frecuencia con la que se combinan. Este número se llama soporte de un elemento o conjunto de elementos. Las correlaciones más fuertes encontradas en el conteo están generando reglas de asociación, cada una con una medida de apoyo, probabilidad e importancia. Para el trabajo investigativo que se realiza no se contempla este tipo de algoritmos, limitándonos a los de Clasificación.

Minería de datos en medicina

Debido a su aplicación exitosa en áreas como el comercio minorista, el comercio electrónico, la banca y los seguros, la minería de datos se ha extendido a otros sectores como la atención médica. La atención médica es un área con un gran potencial para la minería de datos exitosa debido a la gran

cantidad de datos disponibles. Sin embargo, en general todavía hay una falta de herramientas de análisis efectivas para maximizar la utilidad de los datos: a menudo se dice que el entorno de atención médica es rico en información, pero pobre en conocimiento. Se afirman que la minería de datos, y especialmente la minería de datos predictiva, se ha utilizado relativamente en medicina, pero ha comenzado a convertirse en un instrumento importante para los investigadores, pero también para los médicos clínicos. La minería de datos con datos de atención médica difiere de la minería de datos en otras áreas de varias maneras. Estas diferencias pueden causar problemas y requerir más procesamiento previo y preparación de datos que cuando se extraen datos de otras áreas. La primera diferencia se refiere a la calidad de los datos. El objetivo principal de la recopilación de datos de atención médica es mantener un registro del historial médico de un paciente para beneficio de cada paciente. Como resultado, la calidad de los datos a menudo es menor con grandes conjuntos de datos heterogéneos que potencialmente contienen valores perdidos. Como ejemplo, no todos los pacientes, ni siquiera los pacientes con el mismo diagnóstico, se someten a los mismos exámenes y tratamientos que conducen a una amplia variedad de datos. Los registros de salud que no se migran a versiones electrónicas o que solo se escanean pueden contener datos

valiosos para la tarea de minería de datos, pero no se pueden utilizar.

2.3 Definición de términos básicos

a. **Árbol de decisiones. -** Representa un conjunto de reglas de clasificación en forma de un árbol que, a partir de los atributos de cada clase alcanzan un punto final de una ruta.

b. **Asistencia. –** Conjunto de personas que están presentes en un acto.

c. **Clasificación. -** Forma de análisis de datos que permiten extraer modelos que describen las clases importantes de los datos.

d. **IRA**. – Infeccion respiratoria aguda

e. **IRTI. –** Infeccion respiratoria del tracto inferior

f. **IRTS. -** Infeccion respiratoria del tracto superior

g. **Patrón de comportamiento. -** Una forma de conducta que hace las veces de modelo. Los patrones de conducta corresponden a normas específicas, que son guías que orientan la respuesta o acción ante situaciones o circunstancias específicas.

h. **Control.** - Comprobación, inspección, fiscalización, intervención.

i. **Cross-Industry Standard Process for Data Mining (CRISP-DM).** Es una metodología y un modelo de proceso que define un ciclo de seis tareas principales para establecer un marco de trabajo del ciclo de vida de desarrollo en minería de datos.

j. **Dato**. - Información dispuesta de manera adecuada para su tratamiento por una computadora.

k. **Eficiencia**. – Refiere a la optimización de los recursos que se emplea en el proceso, rapidez de respuesta (recurso tiempo) y uso de recursos materiales.

l. **Minería de datos. -** Sistema de información basado en computación que explora grandes repositorios de datos para generar información y descubrir conocimiento.

m. **Modelo. -** Es una representación simplificada de un sistema, construido con el propósito de estudiarlo, donde son considerados los aspectos que afectan al problema de estudio y debe ser lo suficientemente detallado para obtener conclusiones que apliquen al sistema real.

n. **Redes neuronales. -** Modelo de clasificación que consiste en múltiples capas de nodos que simulan el funcionamiento del cerebro humano. Son populares por su gran capacidad de predicción. Son modelos de caja negra por lo que su interpretación puede ser dificultosa.

o. **Sistema. -** Colección de entes que actúan o interactúan para la consecución de un determinado fin. Dados los objetivos del estudio del sistema, generalmente se condiciona el conjunto total de entidades a ser evaluadas.

p. **Support Vector Machines (SVM). -** Algoritmo de clasificación que define hiperplanos basados en los vectores de soporte para crear modelos que permitan establecer la mejor separación entre los puntos por clasificar.

2.4 Formulación de hipótesis

2.4.1 Hipótesis general

El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

2.4.2 Hipótesis especificas

1. Si se identifica los indicadores significativos de enfermedades respiratorias mediante minería de datos entonces se podrá mejorar el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.
2. La aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice significativamente el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco.

2.5 Identificación de variables

Variable Independiente

Reconocimiento de patrones en enfermedades respiratorias

Variables Dependientes

Diagnóstico de pacientes.

2.6 Definición operacional de variables e indicadores

Variable	Indicador
Reconocimiento de patrones en enfermedades respiratorias	▪ Identificación de indicadores en enfermedades respiratorias. ▪ Modelo predictivo de enfermedades respiratorias.
Diagnóstico de pacientes.	▪ Tipo de diagnóstico de pacientes.

CAPITULO III

METODOLOGIA Y TECNICAS DE INVESTIGACIÓN

3.1 Tipo de investigación

La investigación es de tipo aplicada correlacional dado que se intenta establecer un grado de asociación entre las variables bajo estudio.

3.2 Método de investigación

En el desarrollo de la investigación se empleará el método de análisis y síntesis, para ello el trabajo se divide en 2 grupos el grupo muestral de validación para la hipótesis 62 historias clínicas (40% de datos) y, datos de prueba que permitan trabajar con los algoritmos de minería de datos para generar el modelo predictivo, esto es 92 historias clínicas (70% de datos).

3.3 Diseño de la investigación.

Para el estudio se aplicó el diseño de investigación no experimental, ya que no se manipula la variable independiente, y de tipo transeccional porque la información se toma en un solo momento, para su posterior análisis.

3.4 Población y muestra

3.4.1 Población

La población a tomar en cuenta es de 648 historias clínicas de pacientes del hospital Daniel Alcides Carrión en el año 2016, periodo enero - junio.

3.4.2 Muestra

La muestra poblacional la obtenemos a partir de la formula aplicada, siendo de la siguiente forma:

$$n = \frac{Z^2 * p * q * N}{E^2 * N + Z^2 * p * q}$$

Donde:

n: Tamaño de la muestra.

N: Tamaño de la población.

p: porcentaje de la población de acuerdo con la investigación.

q: porcentaje de la población en desacuerdo con la investigación.

Z: Valor Z estadístico, para un nivel de confianza del 90%.

E: error estimado.

Aplicando los siguientes valores con un nivel de confianza del 90%:

N= 648

p= 0.5

q= 0.5

Z=90% que equivale a 1.65

E=10%

Obtenemos n = 62 historias clínicas.

3.5 Técnicas e instrumentos de recolección de datos.

3.5.1 Técnicas.

Para la obtención de los datos e información en la presente investigación se utilizaron:

- La observación.
- El Análisis Bibliográfico.
- TICs

3.5.2 Instrumentos.

- En base a la lectura
- Textos
- Documentos bibliográficos
- Software de aplicación.
- Tabla de Historias clínicas

3.6 Técnica de procesamiento y análisis de datos

Después de hacer la evaluación y crítica de los datos a fin de garantizar la veracidad y confiabilidad se procederá a la depuración de datos innecesarios, mediante las herramientas estadísticas adecuadas, empleando software como Ms Excel y Clementine v 11.1.

3.7 Tratamiento estadístico

A lo largo de los diferentes análisis que se han practicado en este trabajo se ha generado un corpus de datos considerable. Las diferentes variables, tanto cualitativas como cuantitativas, que se han considerado en este trabajo no se pueden concebir como entidades aisladas, sino que deben ser comprendidas dentro de una misma unidad o conjunto de caracteres que forman una globalidad. En este sentido, creemos que es imprescindible el procesamiento de estos datos mediante la aplicación de técnicas estadísticas, ya que su tratamiento sobrepasa ampliamente la capacidad humana.

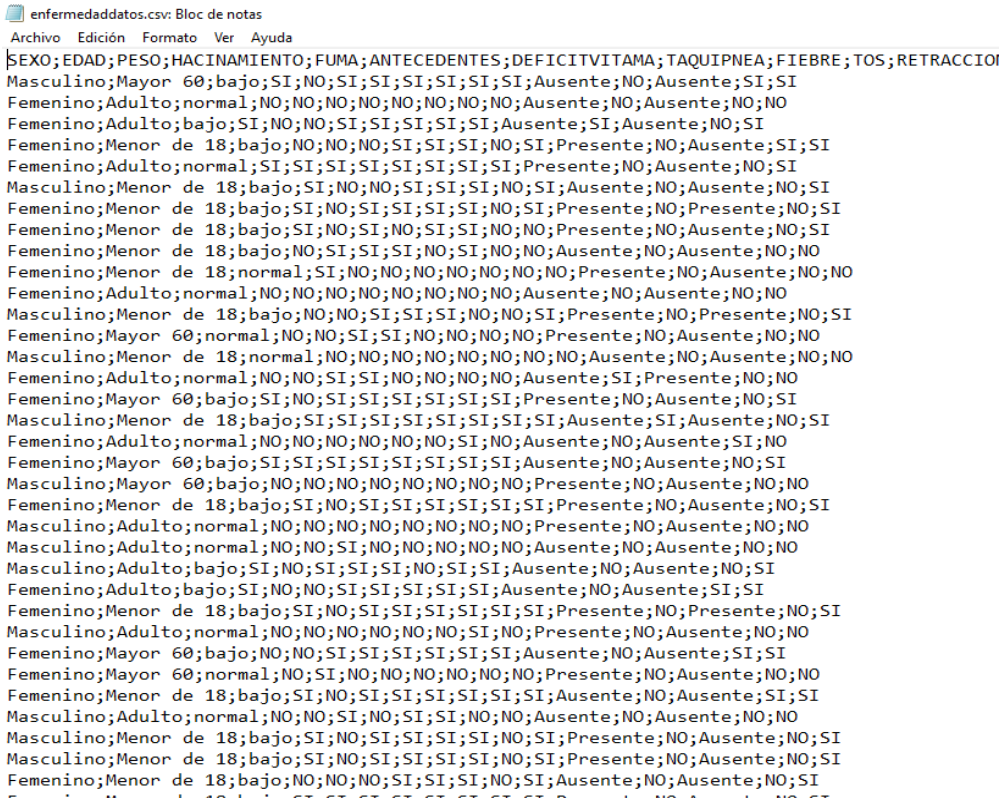
De este modo, el procesamiento estadístico de los datos se revela como un instrumento que se basa en un conjunto de métodos que nos permitirán evidenciar la repartición de los individuos sujetos a estudio en base a los criterios que hemos determinado durante su análisis.

Los recursos necesarios para el procesamiento de los datos y su análisis se han realizado con el programa Clementine v 11.1 y Ms Excel.

3.8 Selección, validación y confiabilidad de los instrumentos de investigación

Selección: Se realizó la selección de 648 historias clínicas de pacientes del Hospital Daniel Alcides Carrión del periodo enero – junio del año 2016 y se tomo la muestra de 242 historias clínicas los cuales se procedió a llevarlos a un formato digital para su procesamiento y análisis.

Validación: Una vez concluido la recolección de los datos; a continuación apoyado en la herramienta de minería de datos pasamos a depurar datos que contienen nulls o espacios vacíos en su correspondiente registro, agrupar y/o categorizar valores que pueden ser muy diversos para una variable predictora, de tal forma que se elimine el ruido al proceder a realizar el modelo de minería de datos



enfermedaddatos.csv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
SEXO;EDAD;PESO;HACINAMIENTO;FUMA;ANTECEDENTES;DEFICITVITAMA;TAQUIPNEA;FIEBRE;TOS;RETRACCIOI
Masculino;Mayor 60;bajo;SI;NO;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Adulto;bajo;SI;NO;NO;SI;SI;SI;SI;SI;Ausente;SI;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;NO;NO;SI;SI;SI;NO;SI;Presente;NO;Ausente;SI;SI
Femenino;Adulto;normal;SI;SI;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;NO;NO;SI;SI;SI;NO;SI;Ausente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Presente;NO;SI
Femenino;Menor de 18;bajo;SI;NO;SI;NO;SI;SI;NO;NO;Presente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;SI;SI;SI;NO;SI;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Menor de 18;normal;SI;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Menor de 18;bajo;NO;NO;SI;SI;SI;NO;NO;SI;Presente;NO;Presente;NO;SI
Femenino;Mayor 60;normal;NO;NO;SI;SI;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Masculino;Menor de 18;normal;NO;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Adulto;normal;NO;NO;SI;SI;NO;NO;NO;NO;Ausente;SI;Presente;NO;NO
Femenino;Mayor 60;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;SI;SI;SI;SI;SI;SI;SI;Ausente;SI;Ausente;NO;SI
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;SI;NO;Ausente;NO;Ausente;SI;NO
Femenino;Mayor 60;bajo;SI;SI;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;NO;SI
Masculino;Mayor 60;bajo;NO;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Masculino;Adulto;normal;NO;NO;SI;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Adulto;bajo;SI;NO;SI;SI;SI;NO;SI;SI;Ausente;NO;Ausente;NO;SI
Femenino;Adulto;bajo;SI;NO;NO;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Presente;NO;SI
Masculino;Adulto;normal;NO;NO;NO;NO;NO;SI;NO;Presente;NO;Ausente;NO;NO
Femenino;Mayor 60;bajo;NO;NO;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Mayor 60;normal;NO;SI;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Masculino;Adulto;normal;NO;NO;SI;NO;SI;SI;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;NO;NO;SI;SI;SI;NO;SI;Ausente;NO;Ausente;NO;SI
```

a. Eliminación de inconsistencias y ruidos de la data

Luego se procede a dar prioridad a los datos que integran el modelo a construir, para ello se elabora una tabla de hechos que permite tener la data origen de las variables predictoras que servirán para generar el modelo predictivo, observe la imagen b.

	SEXO	EDAD	PESO	HACINAMIENTO	FUMA	ANTECEDENTES	DEFICITVITAMA	TAQUIPNEA	FIEBRE	TOS	RETRACCION	ALETEONASAL
2	Masculino	Mayor 60	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Ausente
3	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
4	Femenino	Adulto	bajo	SI	NO	NO	SI	SI	SI	SI	SI	Ausente
5	Femenino	Menor de 18	bajo	NO	NO	NO	SI	SI	SI	NO	SI	Presente
6	Femenino	Adulto	normal	SI	SI	SI	SI	SI	SI	SI	SI	Presente
7	Masculino	Menor de 18	bajo	SI	NO	NO	SI	SI	SI	NO	SI	Ausente
8	Femenino	Menor de 18	bajo	SI	NO	SI	SI	SI	SI	NO	SI	Presente
9	Femenino	Menor de 18	bajo	SI	NO	SI	NO	SI	SI	NO	NO	Presente
10	Femenino	Menor de 18	bajo	NO	SI	SI	SI	NO	SI	NO	NO	Ausente
11	Femenino	Menor de 18	normal	SI	NO	NO	NO	NO	NO	NO	NO	Presente
12	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
13	Masculino	Menor de 18	bajo	NO	NO	SI	SI	SI	NO	NO	SI	Presente
14	Femenino	Mayor 60	normal	NO	NO	SI	SI	NO	NO	NO	NO	Presente
15	Masculino	Menor de 18	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
16	Femenino	Adulto	normal	NO	NO	SI	SI	NO	NO	NO	NO	Ausente
17	Femenino	Mayor 60	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Presente
18	Masculino	Menor de 18	bajo	SI	SI	SI	SI	SI	SI	SI	SI	Ausente
19	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	SI	NO	Ausente
20	Femenino	Mayor 60	bajo	SI	SI	SI	SI	SI	SI	SI	SI	Ausente
21	Masculino	Mayor 60	bajo	NO	NO	NO	NO	NO	NO	NO	NO	Presente
22	Femenino	Menor de 18	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Presente
23	Masculino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Presente

b.Hechos con parte de los datos válidos, según las variables predictoras.

3.9 Orientación ética

El presente trabajo bajo los factores morales y principios éticos en la investigación, de acuerdo a los criterios establecidos en las guías, así como el planeamiento y fundamentación establecidos para el desarrollo de la investigación.

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1 Descripción del trabajo de campo

Se hizo un estudio aplicado de tipo correlacional, no experimental, se vincularon al estudio 92 historias clínicas del hospital Daniel Alcides Carrión que se atendieron por temas de enfermedades respiratorias, todos mayores de edad, pertenecientes al año 2016 en el periodo enero – julio, la información contenida en los formatos físicos, fue posteriormente interpretada y transferida a una base de datos creada con una estructura acorde a las necesidades de minería de datos.

Agenciada la información se procedió a limpiar los campos que no contenían data con la finalidad de eliminar el ruido para evitar que el

modelo de minería de datos arroje resultados erróneos o con sesgos pronunciados.

La información de las historias clínicas se distribuyó en los 15 descriptores predictivos para su análisis (indicadores de neumonía entre síntomas y signos) y un descriptor resultado diagnostico cuyos valores posibles fueron 2: presenta neumonía y no presenta neumonía.

Los datos para la mayoría de los indicadores predictores son cualitativos. Así mismo se considera los aspectos éticos de no divulgar información sensible de las historias clínicas más allá de lo que la investigación requiera.

El modelo de minería de datos fue construido y luego probado con el 60% de los datos, esto es 92 historias clínicas, quedando el 40% restante para realizar pruebas de validez del modelo creado, 62 datos muestrales

4.1.1 Diagnostico organizacional del Hospital Daniel Alcides Carrión - Pasco

El Hospital Daniel Alcides Carrión – Pasco es una institución del estado creada para garantizar una atención integral de salud ajustándose a los lineamientos de política sectorial del Ministerio de salud los cuales son:

- Promoción de Salud y prevención de Enfermedad.
- Extensión y universalización del aseguramiento en salud.
- Desarrollo de los recursos humanos con respeto y dignidad

- Creación del sistema coordinado y Descentralizado de Salud.
- Modernización del MINSA y fortalecimiento en su rol de conducción social.
- Nuevo modelo de atención integral de salud.
- Democratización de salud.
- Financiamiento interno y externo orientado a los sectores más pobres.

El Hospital Daniel A. Carrión trata de optimiza sus servicios en relación con su zona de influencia, estableciendo Centros y Puestos de Salud en su radio de influencia que cubre el primer nivel de atención.

Figura 4.1. Vista aérea del Hospital Daniel A. Carrión

4.1.2 Ubicación Geográfica

El hospital Daniel Alcides Carrión se encuentra ubicado en el distrito de Yanacancha, ciudad de San Juan Pampa, Av. Daniel A. Carrión s/n y con un segundo acceso por la Av. Los Incas s/n. (figura 4.2). Sin embargo, actualmente viene operando de forma momentánea en la zona de la Esperanza frente a las oficinas administrativas de EsSalud Pasco (figura 4.3), conocida como Casa de Piedra, debido a la remodelación total que se esta realizando con una nueva infraestructura que brinde mejores prestaciones.

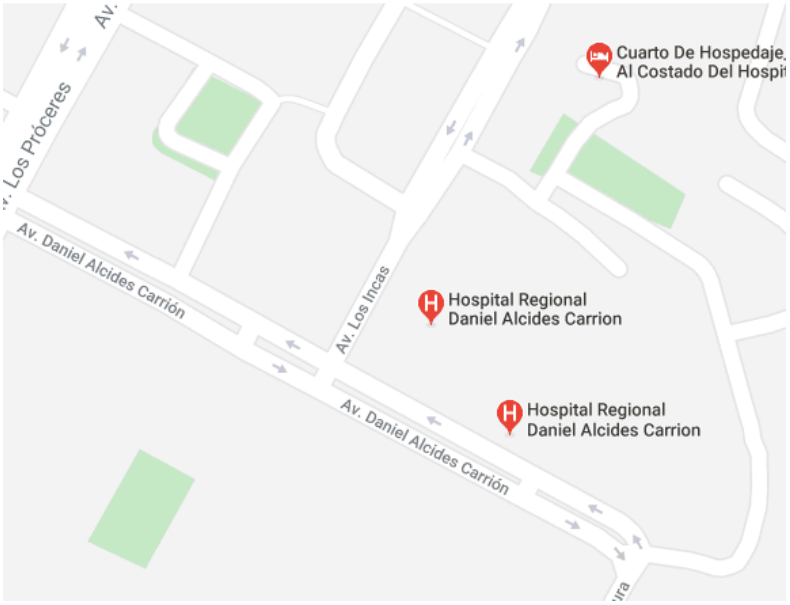


Figura 4.2 Ubicación geográfica del Hospital Daniel A. Carrión.

Figura 4.3 Infraestructura momentánea del Hospital Daniel A. Carrión.

4.1.3 Misión

El Hospital Daniel Alcides Carrión en una institución de salud organizada que brinda y garantiza una atención integral de salud de las personas, familias y la población en general, priorizando a la población más necesitada de Pasco.

4.1.4 Visión

Ser una institución de la salud líder en la comunidad con personal calificado con tecnología e infraestructura moderna y adecuada que garantice lograr el bienestar de salud en la población, ofertando servicios especializados con la más alta calidad y calidez.

4.1.5 Estructura organizacional

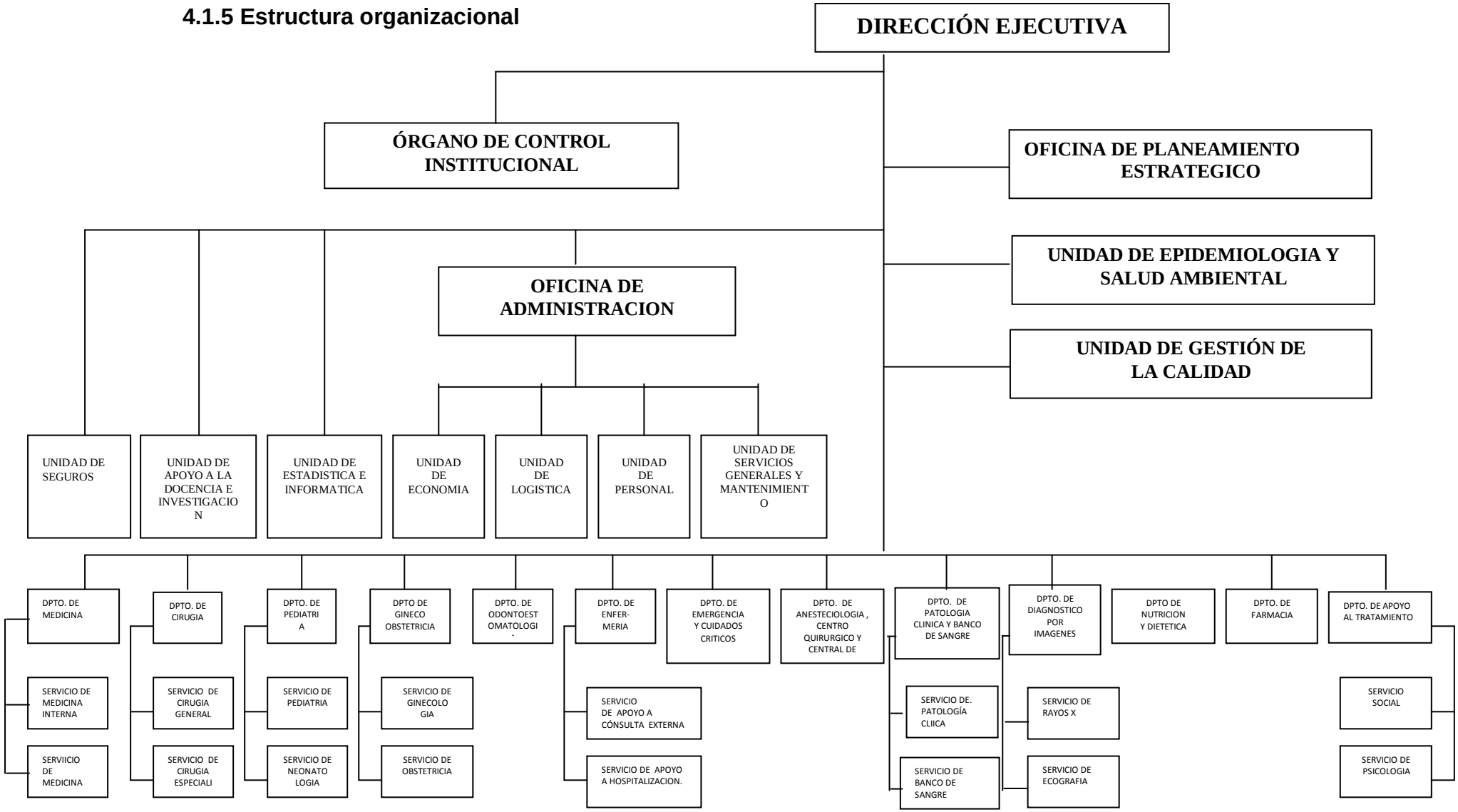


Figura 4.4. Organigrama del Hospital Daniel A. Carrió

4.1.6 Enfermedades respiratorias en Pasco.

Las enfermedades respiratorias o infecciones respiratorias agudas (IRA), conforme se indicó en las bases teóricas de esta investigación, son las enfermedades infecciosas más frecuentes en el ser humano. La población más vulnerable a nivel nacional y por consiguiente en la región Pasco son los niños de 0 a 5 años y los adultos mayores desde 60 años en adelante. Según estadísticas del Ministerio de Salud los niños pueden presentar entre cinco y ocho infecciones respiratorias al año, acentuándose mas en zonas de clima frígido como es la región Pasco. Así mismo el Ministerio de Salud indica que las infecciones respiratorias bajas (neumonías) continúan siendo la primera causa de muerte en el Perú (vea figura 4.4), esto según un estudio del 2016.

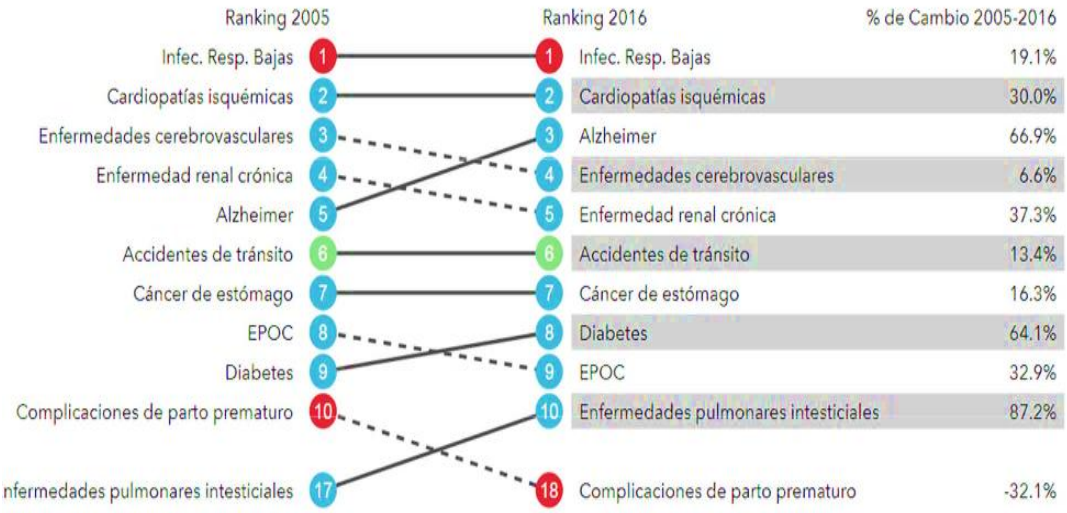


Figura 4.4. Diez principales causas de muerte por enfermedad en el Perú.

El mismo estudio indico que en el año 2017 las infecciones respiratorias agudas presentes en la región Pasco oscilaron entre el 20 y 24% del total registrado en el país, conforme se aprecia en la figura 4.5.

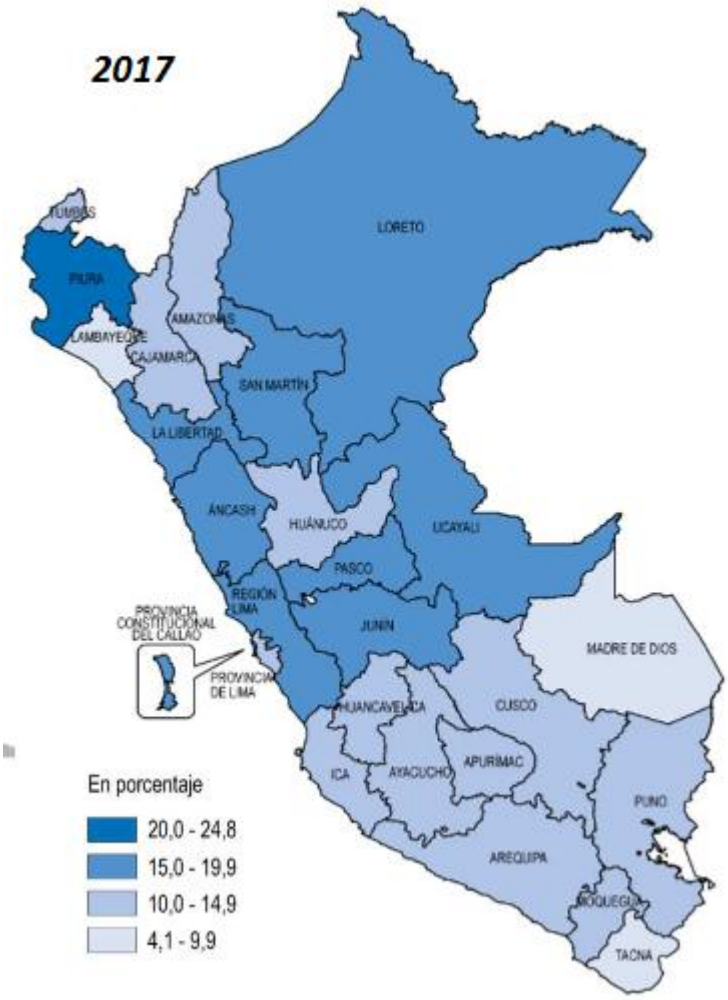


Figura 4.5. Rango porcentual de enfermedades respiratorias en el Perú

Por otro lado, el número de defunciones presentados en la región Pasco por temas de neumonía ha ido disminuyendo entre el 2015 y 2018, conforme se puede apreciar en la tabla 4.1.

Departamento	Menores de 5 años				Def. < 5 años				Mayores de 60 años				Def. > 60 años			
	2015	2016	2017	2018	2015	2016	2017	2018	2015	2016	2017	2018	2015	2016	2017	2018
Amazonas	233	207	246	217	5	4	6	3	74	63	115	100	0	3	3	2
Ancash	298	251	418	346	2	3	0	1	128	91	223	299	1	5	8	4
Apurimac	207	149	167	167	1	0	2	1	74	97	105	115	2	4	6	4
Arequipa	754	699	894	627	0	2	1	3	527	737	718	755	13	19	31	21
Ayacucho	340	105	184	164	5	3	2	4	155	102	128	144	8	21	16	20
Cajamarca	477	312	382	282	2	3	2	2	45	93	234	384	4	3	5	9
Callao	522	482	414	389	1	2	1	2	265	391	380	363	18	31	51	46
Cusco	517	289	453	511	18	8	9	17	297	382	458	618	12	24	21	39
Huancavelica	201	110	114	112	8	3	11	3	126	133	109	141	29	15	15	14
Huanuco	575	516	470	522	1	5	5	9	183	215	207	151	8	5	10	7
Ica	231	184	178	179	0	1	1	0	77	66	54	60	24	25	34	36
Junin	460	146	286	289	10	5	10	10	360	196	183	249	11	12	5	14
La Libertad	316	281	273	321	11	8	4	5	302	425	417	486	27	66	120	41
Lambayeque	342	230	146	194	1	3	3	0	26	53	25	69	0	0	2	1
Lima	3860	4691	5373	4857	8	22	13	4	1271	1834	2124	2534	119	293	289	166
Loreto	1324	1043	856	946	13	16	12	16	164	264	113	209	5	3	6	4
Madre de Dios	84	106	164	135	1	1	2	0	25	10	15	41	1	0	0	1
Moquegua	45	23	73	48	0	0	2	1	50	54	88	100	1	0	4	15
Pasco	207	142	191	173	7	3	2	2	76	46	52	76	5	2	2	0
Piura	921	695	758	585	4	2	3	2	273	607	891	794	14	26	28	22
Puno	416	439	507	410	13	11	16	8	303	310	369	432	23	17	14	14
San Martin	337	195	244	186	3	4	1	0	74	61	92	74	6	1	2	0
Tacna	27	16	25	38	0	0	4	1	10	18	4	29	3	1	0	2
Tumbes	77	80	131	50	0	0	0	0	36	28	48	68	2	1	1	0
Ucayali	686	781	645	595	2	5	8	8	48	65	114	215	1	0	0	2
Total	13457	12172	13592	12263	116	114	120	102	4969	6341	7266	8506	337	577	673	484

Fuente: Centro Nacional de Epidemiología, Prevención y Control de Enfermedades – MINSA
Elaborado por CDC

Tabla 4.5. Número de casos presentados y de defunciones en el país por neumonía.

Con todos estos antecedentes se optó por focalizar la investigación en el diagnóstico de las neumonías por ser la principal causa de muerte dentro de las enfermedades y en particular de las enfermedades respiratorias y además por presentar las mayores tasas de presencia a nivel nacional dentro de la región Pasco.

Aplicación de minería de datos

Formulación del problema

La neumonía es una enfermedad común que sigue siendo la principal causa de muerte de niños pequeños en países en desarrollo y personas mayores en países desarrollados.

La neumonía no es más que un proceso inflamatorio agudo del parénquima pulmonar, de origen infeccioso, que se inicia fuera del ambiente hospitalario. Se caracteriza por la aparición de fiebre y/o síntomas respiratorios, junto con la presencia de infiltrados pulmonares en la radiografía de tórax.

La clasificación de la neumonía se da entres grandes grupos: Neumonía típica causada por bacterias, neumonía atípica causada por virus o bacterias atípicas y finalmente los no clasificables que no se pueden incluir en ninguna de las 2 anteriores.

Preprocesamiento de datos

Síntomas y signos de la Neumonía. La neumonía presenta características básicas que se distribuyen entre síntomas y signos, estos se mencionan a continuación:

- **Taquipnea:** es el síntoma con mayor sensibilidad para el diagnóstico de neumonía comparado con la radiografía de tórax. un aumento de la frecuencia respiratoria por

encima de los valores normales (>20 inspiraciones por minuto). Se catalogo como Si y No, es decir si presenta este síntoma o no presenta el síntoma.

- **Fiebre:** generalmente es súbita, mayor de 38,5°C, asociada con frecuencia a escalofríos en las infecciones bacterianas. Se catalogo en Si y No, donde si indica una temperatura elevada igual o mayor a 38.5° C y, no cuando su temperatura es por debajo de este nivel.

- **Tos:** Casi siempre es seca al inicio del padecimiento; posteriormente, húmeda, acompañada de expectoración en los niños mayores de ocho años, ya que antes de esta edad no es posible. Se catalogo como Si presenta tos y No presenta tos.

- **Retracción:** Moderada a severa y dificultad respiratoria grave. Se catalogó como Si presenta retracción o No presenta retracción.

- **Aleteo nasal:** cuando las fosas nasales se ensanchan cuando se respira. Se catalogo como presente o ausente.

- **Cianosis o hipoxemiasaturación:** Coloración azul o lívida de la piel y de las mucosas que se produce a causa de una oxigenación deficiente de la sangre. Menor del 88% a la altura de Pasco o menor del 92% a nivel del

mar. Se catalogo con Si cuando se dio el cuadro y No cuando no se observó tal situación.

- **Apnea:** Colapso en las vías respiratorias o una obstrucción de ellas durante el sueño. Luego, la respiración vuelve con un ronquido o resoplido. Se catalogo presente o ausente.

- Signos de **deshidratación**, se catalogó como Si y No.

- **Edad:** La edad se categorizo en tres niveles: menores de 18 años, adultos y mayores de 60 años.

- **Peso:** La falta de peso es otro signo de riesgo. Se categoriza en normal y bajo de peso.

- **Hacinamiento:** del lugar donde reside, se tomó como hacinado si viven en máximo dos ambientes (cuartos) y no si la casa posee más de dos ambientes, se catalogó como Si y No.

- **Fumar:** Un factor de riesgo presente para contraer neumonía. Se catalogo entre fumador y no fumador (Si y No).

- **Déficit de vitamina A.** Catalogado como Si y No.

- **Sexo:** El sexo se categoriza en masculino y femenino.

- **Antecedentes:** Haber contraído la enfermedad con anterioridad. Se catalogaron en Si y No.

Estas 15 características a las que se denominaran indicadores de neumonía son las que se tomaron en consideración para ser evaluadas y determinar el grado de incidencia al momento de pronosticar o predecir en el modelo predictivo mediante minería de datos el **diagnóstico de neumonía** como enfermedad respiratoria que es la variable dependiente o variable predecida.

Construcción de la base de datos para minería de datos. Basados en los indicadores (síntomas y signos de la neumonía) se procedió a construir la base de datos y alimentarla con la información de las historias clínicas, 92 en total, del periodo 2016, que fueron proporcionadas de manera confidencial y con el compromiso ético de no revelar la información personal de los pacientes, y de no dar ningún uso más que el indicado en la investigación.

La construcción de la base de datos presenta la siguiente forma:

INDICADORES PREDICTORES → DIAGNOSTICO

Los indicadores predictores corresponden a las 15 características descritas en el ítem 4.1.6.2, y el indicador a

predecir es el Diagnostico que presenta solo 2 valores: Presenta neumonía / no presenta neumonía (SI o NO).

Basados en esta información como punto de inicio se construye la base de datos que se muestra en la figura 4.6, la que almaceno la información de las historias clínicas de los pacientes que fueron atendidos en el hospital Daniel Alcides Carrión con posibles síntomas de enfermedades respiratorias.

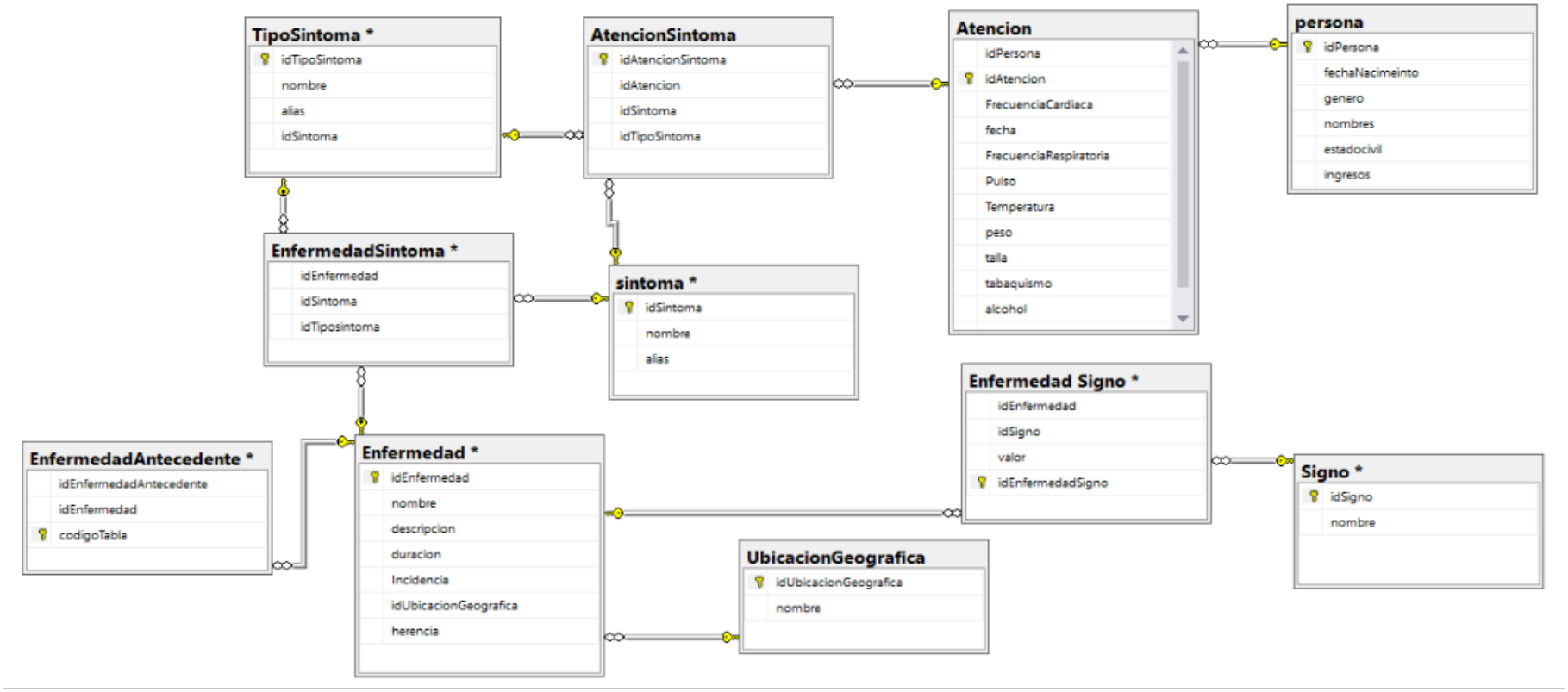


Figura 4.6. Estructura de la base de datos para el proceso de minería de datos.

La base de datos tendrá en promedio un tamaño de 10 Mb escalable a un mayor tamaño en función de la data que se almacene para robustecer el trabajo en minería de datos, la descripción de cada tabla que conforma la base de datos se indica en la tabla 4.2.

Descripción de las tablas a usar

TABLA	DESCRIPCION
Enfermedad	Enfermedades Respiratorias
Signo	Signos de la enfermedad
Sintoma	Síntomas de la enfermedad
Enfermedad Antecedente	Antecedente de enfermedades
Tipo Síntoma	Tipo de Síntoma
Persona	Paciente de un centro médico
Atencion	Atenciones generadas por la persona

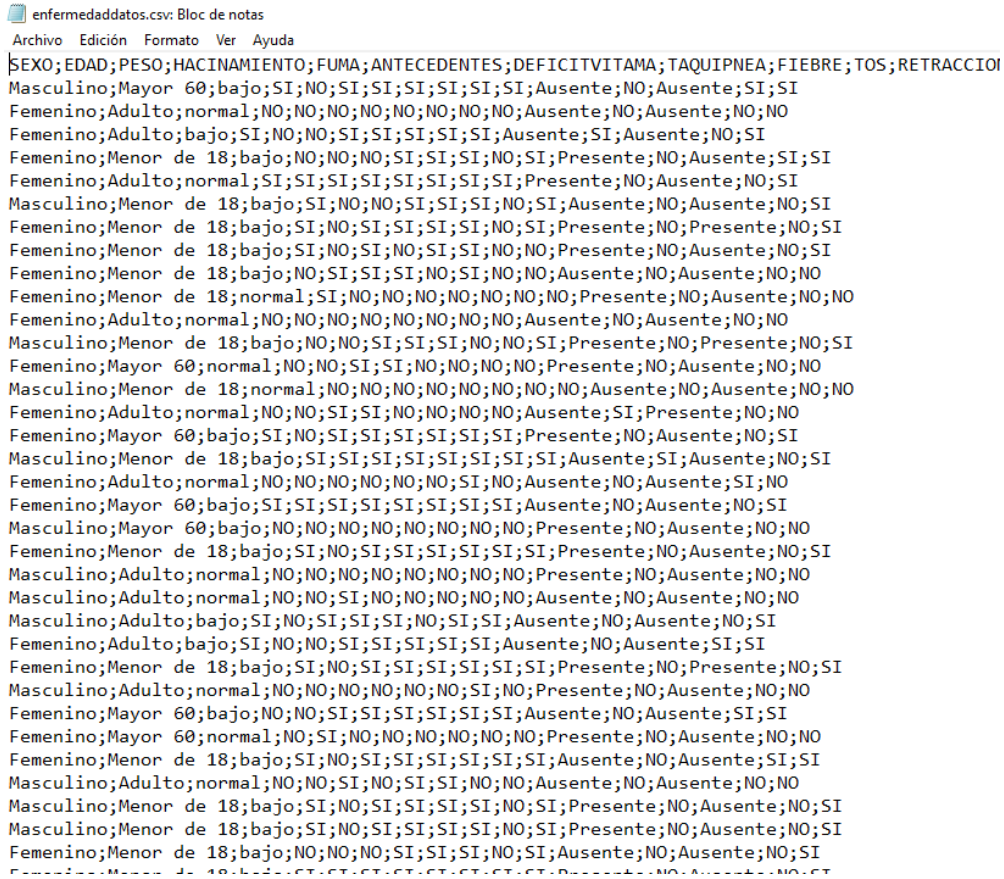
Tabla 4.2. Descripción de las tablas de la base de datos.

Luego se procede a dar prioridad a los datos que integran el modelo a construir, para ello se elabora una tabla de hechos que permite tener la data origen de las variables predictoras que servirán para generar el modelo predictivo, observe la figura 4.7.

	SEXO	EDAD	PESO	HACINAMIENTO	FUMA	ANTECEDENTES	DEFICITVITAMA	TAQUIPNEA	FIEBRE	TOS	RETRACCION	ALETEONASAL
2	Masculino	Mayor 60	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Ausente
3	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
4	Femenino	Adulto	bajo	SI	NO	NO	SI	SI	SI	SI	SI	Ausente
5	Femenino	Menor de 18	bajo	NO	NO	NO	SI	SI	SI	NO	SI	Presente
6	Femenino	Adulto	normal	SI	SI	SI	SI	SI	SI	SI	SI	Presente
7	Masculino	Menor de 18	bajo	SI	NO	NO	SI	SI	SI	NO	SI	Ausente
8	Femenino	Menor de 18	bajo	SI	NO	SI	SI	SI	SI	NO	SI	Presente
9	Femenino	Menor de 18	bajo	SI	NO	SI	NO	SI	SI	NO	NO	Presente
10	Femenino	Menor de 18	bajo	NO	SI	SI	SI	NO	SI	NO	NO	Ausente
11	Femenino	Menor de 18	normal	SI	NO	NO	NO	NO	NO	NO	NO	Presente
12	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
13	Masculino	Menor de 18	bajo	NO	NO	SI	SI	SI	NO	NO	SI	Presente
14	Femenino	Mayor 60	normal	NO	NO	SI	SI	NO	NO	NO	NO	Presente
15	Masculino	Menor de 18	normal	NO	NO	NO	NO	NO	NO	NO	NO	Ausente
16	Femenino	Adulto	normal	NO	NO	SI	SI	NO	NO	NO	NO	Ausente
17	Femenino	Mayor 60	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Presente
18	Masculino	Menor de 18	bajo	SI	SI	SI	SI	SI	SI	SI	SI	Ausente
19	Femenino	Adulto	normal	NO	NO	NO	NO	NO	NO	SI	NO	Ausente
20	Femenino	Mayor 60	bajo	SI	SI	SI	SI	SI	SI	SI	SI	Ausente
21	Masculino	Mayor 60	bajo	NO	NO	NO	NO	NO	NO	NO	NO	Presente
22	Femenino	Menor de 18	bajo	SI	NO	SI	SI	SI	SI	SI	SI	Presente
23	Masculino	Adulto	normal	NO	NO	NO	NO	NO	NO	NO	NO	Presente

Figura 4.7. Tabla de hechos con parte de los datos válidos, según las variables predictoras.

A continuación apoyado en la herramienta de minería de datos pasamos a depurar datos que contienen nulls o espacios vacíos en su correspondiente registro, agrupar y/o categorizar valores que pueden ser muy diversos para una variable predictora, de tal forma que se elimine el ruido al proceder a realizar el modelo de minería de datos, conforme se observa en la figura 4.8, donde se observa los datos depurados en un archivo denominado enfermedaddatos.csv, este archivo y con este formato será el que se emplee para realizar el análisis predictivo en el software Clementin v11.1.


enfermedaddatos.csv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
SEXO;EDAD;PESO;HACINAMIENTO;FUMA;ANTECEDENTES;DEFICITVITAMA;TAQUIPNEA;FIEBRE;TOS;RETRACCIOI
Masculino;Mayor 60;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Adulto;bajo;SI;NO;NO;SI;SI;SI;SI;SI;Ausente;SI;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;NO;NO;SI;SI;SI;NO;SI;Presente;NO;Ausente;SI;SI
Femenino;Adulto;normal;SI;SI;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;NO;NO;SI;SI;SI;NO;SI;Ausente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Presente;NO;SI
Femenino;Menor de 18;bajo;SI;NO;SI;NO;SI;SI;NO;NO;Presente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;SI;SI;SI;NO;SI;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Menor de 18;normal;SI;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Menor de 18;bajo;NO;NO;SI;SI;SI;NO;NO;SI;Presente;NO;Presente;NO;SI
Femenino;Mayor 60;normal;NO;NO;SI;SI;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Masculino;Menor de 18;normal;NO;NO;NO;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Femenino;Adulto;normal;NO;NO;SI;SI;NO;NO;NO;NO;Ausente;SI;Presente;NO;NO
Femenino;Mayor 60;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;SI;SI;SI;SI;SI;SI;SI;Ausente;SI;Ausente;NO;SI
Femenino;Adulto;normal;NO;NO;NO;NO;NO;NO;SI;NO;Ausente;NO;Ausente;SI;NO
Femenino;Mayor 60;bajo;SI;SI;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;NO;SI
Masculino;Mayor 60;bajo;NO;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Ausente;NO;SI
Masculino;Adulto;normal;NO;NO;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Masculino;Adulto;normal;NO;NO;SI;NO;NO;NO;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Adulto;bajo;SI;NO;SI;SI;SI;NO;SI;SI;Ausente;NO;Ausente;NO;SI
Femenino;Adulto;bajo;SI;NO;NO;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Presente;NO;Presente;NO;SI
Masculino;Adulto;normal;NO;NO;NO;NO;NO;NO;SI;NO;Presente;NO;Ausente;NO;NO
Femenino;Mayor 60;bajo;NO;NO;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Femenino;Mayor 60;normal;NO;SI;NO;NO;NO;NO;NO;NO;Presente;NO;Ausente;NO;NO
Femenino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;SI;SI;Ausente;NO;Ausente;SI;SI
Masculino;Adulto;normal;NO;NO;SI;NO;SI;SI;NO;NO;Ausente;NO;Ausente;NO;NO
Masculino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Ausente;NO;SI
Masculino;Menor de 18;bajo;SI;NO;SI;SI;SI;SI;NO;SI;Presente;NO;Ausente;NO;SI
Femenino;Menor de 18;bajo;NO;NO;NO;SI;SI;SI;NO;SI;Ausente;NO;Ausente;NO;SI
```

Figura 4.8. Eliminación de inconsistencias y ruidos de la data.

A continuación, se procede a trabajar en el Clementin versión 11.1, importando la Tabla de hechos ya creada y optimizada de nombre enfermedaddatos.csv, conforme se aprecia en la figura 4.9

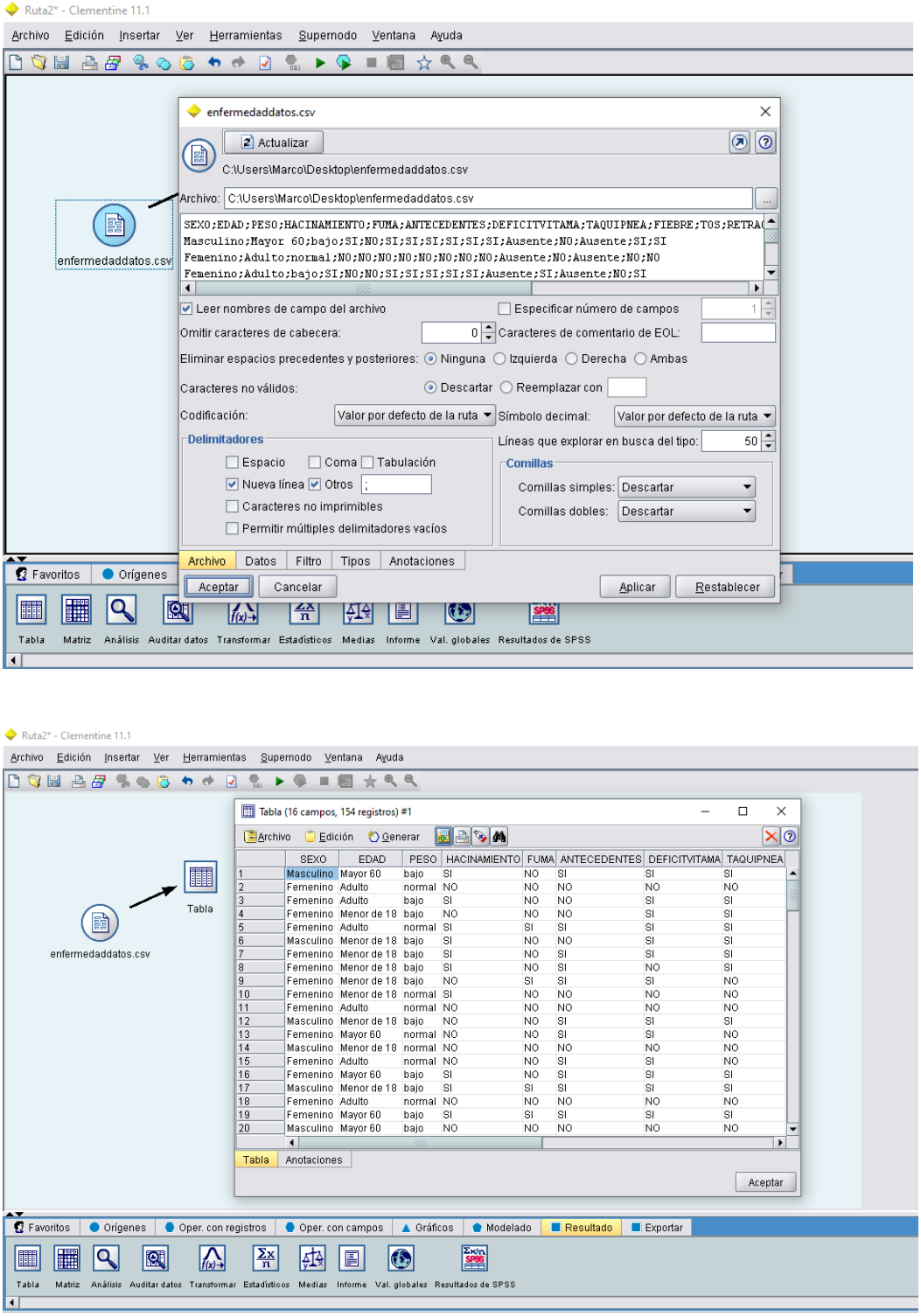


Figura 4.9. Importación de la tabla de hechos a Clementin.

Una vez importada la data se procede a seleccionar los elementos o variables predictoras, que se catalogara como entradas así como la variable resultado o salida. Vea figura 4.10.

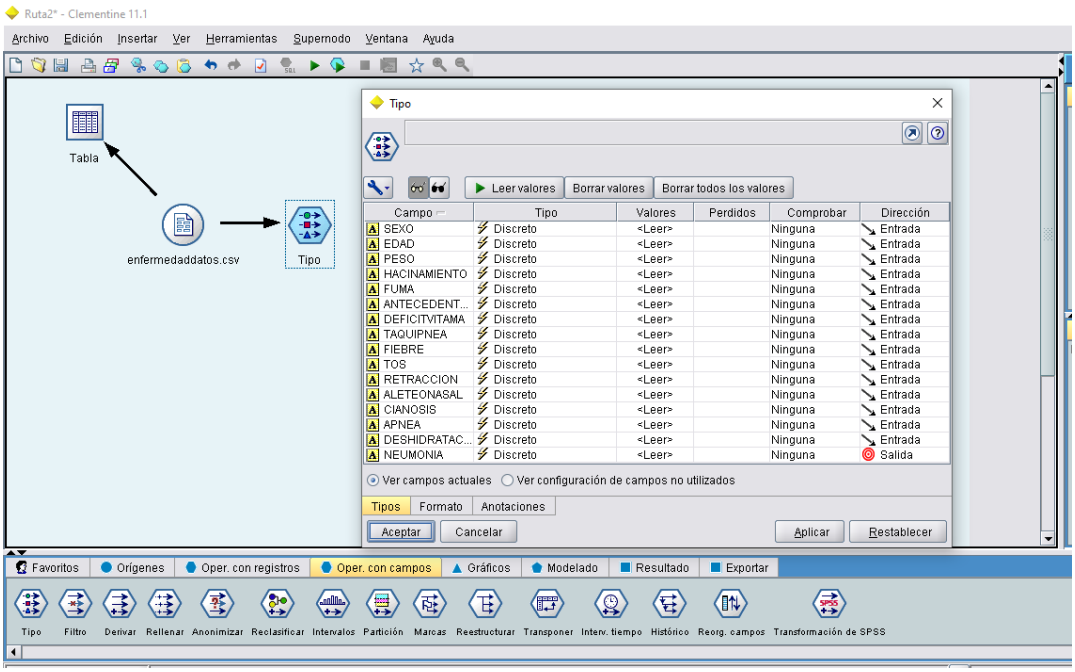


Figura 4.10. Tipos de variables declaradas en Clementin

Mencionar que la variable salida es NEUMONÍA mientras que las demás (15 variables predictoras) son de tipo entrada.

Modelo de construcción

A continuación, se seleccionó la técnica de modelado, se optó por el modelo de clasificación mediante la prueba de tres algoritmos:

- **Árbol de clasificación y regresión (C&R)** genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

- **El nodo C5.0,** genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

- **El nodo CHAID,** genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&R y C5.0, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

Basados en estos tres algoritmos de árboles de decisión se construye los modelos predictivos de la investigación, figura 4.11, mediante el entrenamiento con los 92 registros de datos, obteniéndose los siguientes niveles predictivos de diagnóstico de neumonía en función de las variables descriptoras:

Algoritmo árbol de clasificación y regresión (C&R)

Correctos	92	100%
Erróneos	0	0%
Total	92	

Algoritmo Nodo C5.0

Correctos	88	95,65%
Erróneos	4	4,35%
Total	92	

Algoritmo nodo CHAID

Correctos	91	98,91%
Erróneos	1	1,09%
Total	92	

Como se puede apreciar el algoritmo CyR resulta siendo el mas eficiente en el entrenamiento, con un 100% de éxito predictivo en todos los casos de diagnostico de neumonía, por lo que se considera el mas apropiado para esta investigación, sin embargo los otros dos algoritmos también presentan niveles predictivos muy altos por lo que pondrán a prueba los tres algoritmos para la validación de las hipótesis.

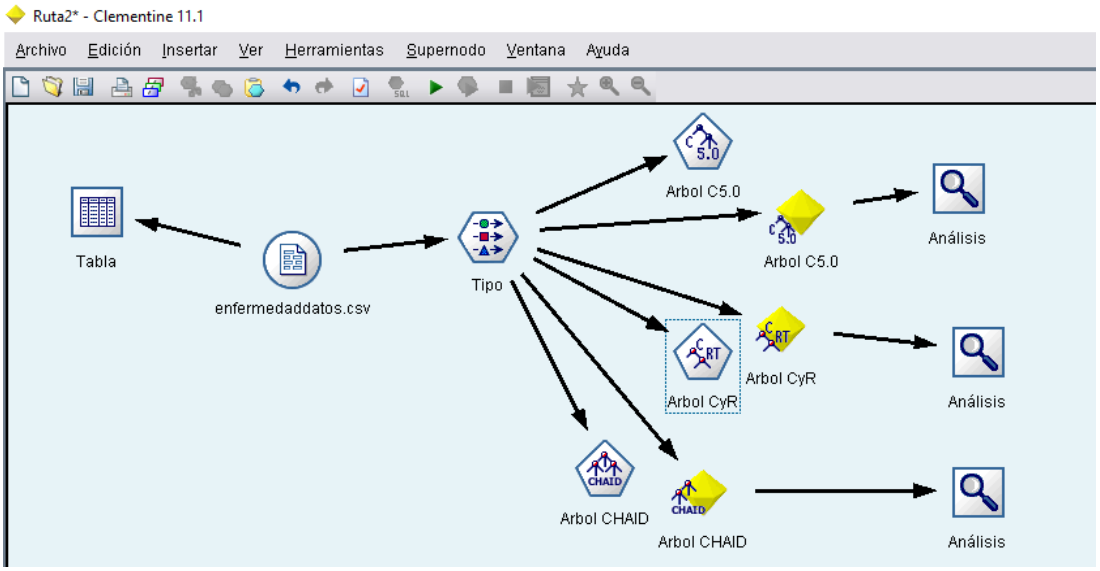


Figura 4.11. Aplicación de algoritmos de árboles de decisión.

Encontrado el modelo, en el capítulo V se evalúa con los datos de la muestra de estudio (62 registros) para validar las hipótesis de la investigación.

4.2 Presentación, análisis e interpretación de resultados

A continuación, se presentan los resultados obtenidos de la aplicación de minería de datos de los patrones en enfermedades respiratorias que se sustentan en la investigación de la hipótesis de investigación:

Hg: El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

Para la hipótesis especifica h1, el análisis realizado entre los predictores que influyen significativamente en el diagnóstico, se establece que 5 de ellos del total de 15 (vea tabla 5.1) aportan un valor de significancia superior al 50%, esta información es proporcionada por los resultados del modelo predictivo mediante el algoritmo de árbol de decisión CyR, modelo que se empleó por ser el que mejor resultados dio en los entrenamientos del capítulo anterior, a continuación se muestra el árbol de decisión (figura 5.2) donde se aprecia las variables predictoras que influyen para el diagnóstico de neumonía, empleando el archivo enfermedadmuestra.csv como se aprecia en la figura 5.1 :

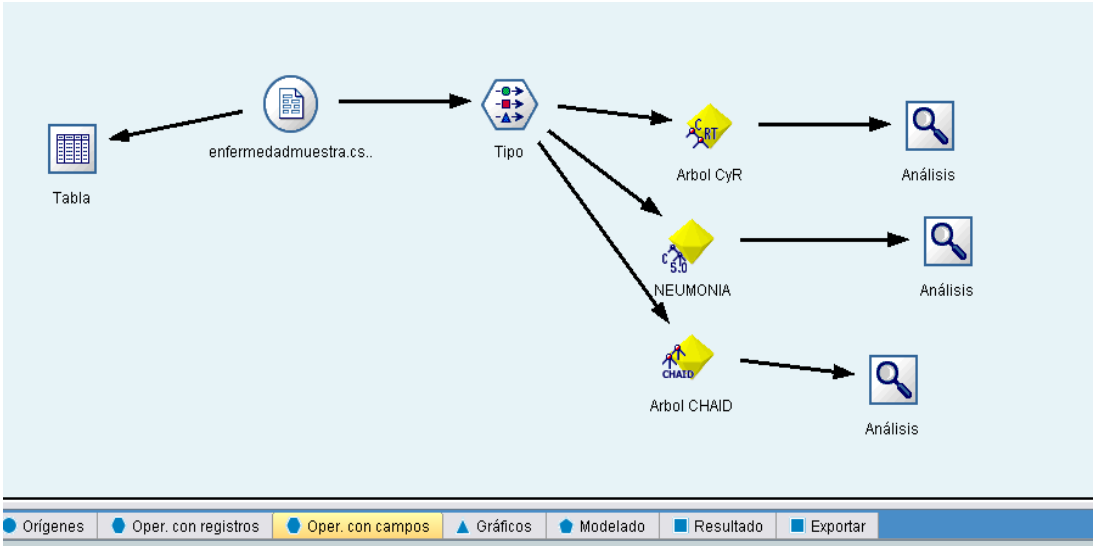


Figura 5.1. Prueba de los modelos predictores con la muestra de estudio.

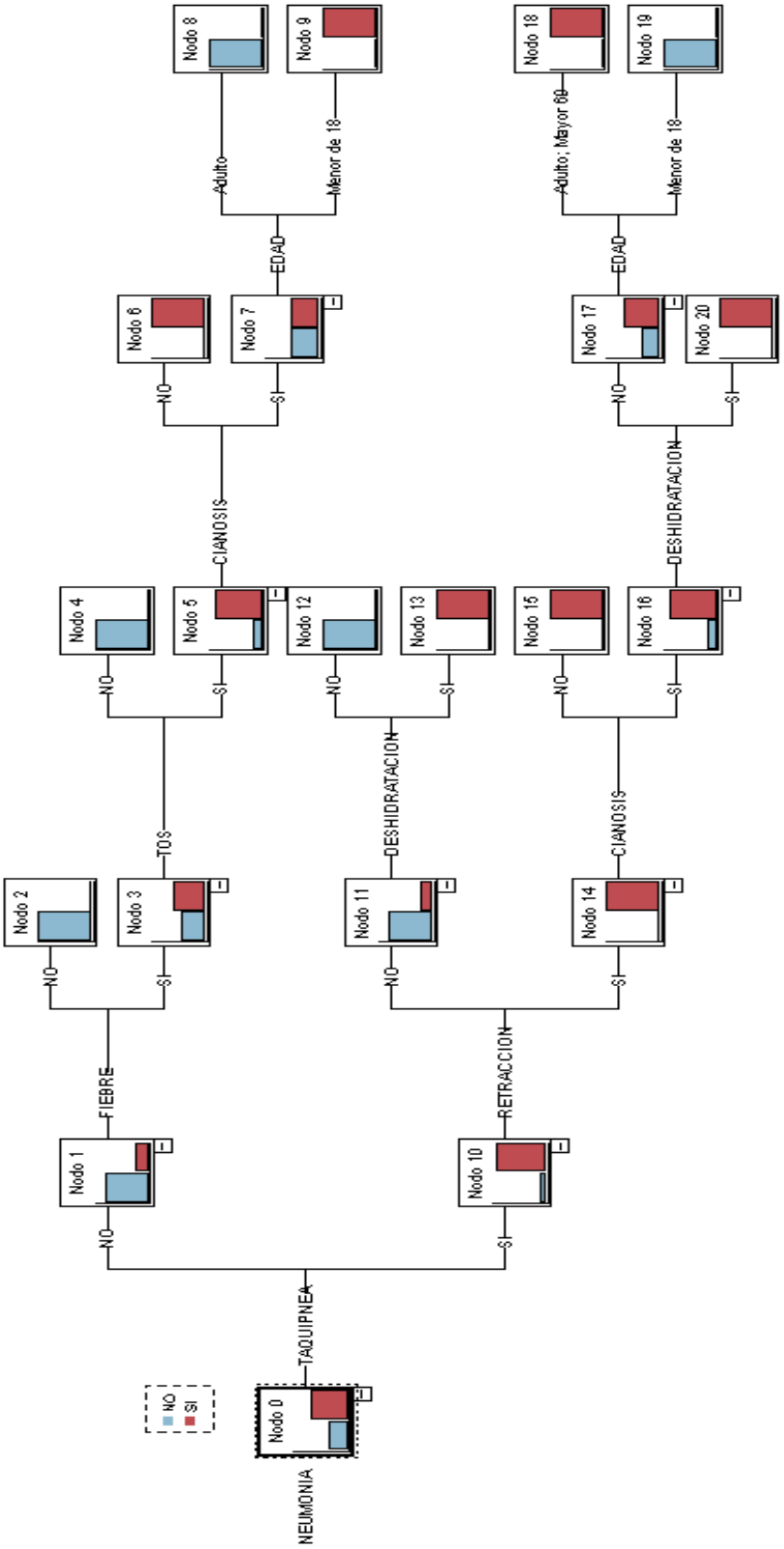


Figura 5.2 Modelo predictivo algoritmo árbol de decisión.

N	Predictor	Nivel de predicción
1	Taquipnea	78%
2	Fiebre	62%
3	Retracción	71%
4	Deshidratación	56%
5	Tos	71%
6	Edad	68%

Tabla 5.1. Nivel de predicción de los indicadores de neumonía.

Esto indica que 6 de los indicadores son significativos y deben ser incluidos en el modelo final predictivo de la neumonía. Por lo que la hipótesis especifica 1:

> H1: Si se identifica los indicadores significativos de enfermedades respiratorias mediante minería de datos entonces se podrá mejorar el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

Queda validada y demostrada en la investigación.

En el caso de la segunda hipótesis específica, la predicción que arrojan las distintas técnicas predictivas de minería de datos para el caso analizado se muestra en la tabla 5.1. Estos resultados manifiestan la existencia de evidencia suficiente para indicar que, si es posible realizar predicciones con un modelo que incluye los indicadores significativos que influyen en las enfermedades respiratorias, como es la neumonía, y que predicen con un alto grado de acierto en el diagnóstico de neumonía. Los valores alcanzados oscilaron en un valor máximo de 100% en el mejor modelo encontrado.

Modelos según algoritmos de clasificación	Predicción
Arbol de clasificación y regresión (C&R)	95.16%
Árbol de decisión C5.0	100%
nodo CHAID	98.39%

Tabla 5.2. Resultados de la predicción con los diferentes algoritmos de minería de datos.

Los diferentes datos encontrados en la información extraída han sido analizados y evaluados en el modelo de minería de datos que arrojaron una predicción acertada de 100% en el mejor escenario empleando el algoritmo de árbol de decisión C5.0, superando al modelo CyR que en el caitulo IV fue el quien arrojo mejores resultados, esto se puede atribuir a la cantidad de data evaluada, es decir el algoritmo C5.0 brinda mejores resultados ante tamaños de muestra menores, sin embargo el nivel de predicción de los 3 algoritmos es muy alto.

Esto permite indicar que el modelo predictivo sobre enfermedades respiratorias resulta significativo, por lo que la hipótesis especifica:

H2: La aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice significativamente el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco.

Resulta validada en la investigación.

Finalmente, **la hipótesis general** de la investigación “El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco”, se considera aceptada como consecuencia de la aceptación de las hipótesis específicas.

4.3 Prueba de hipótesis

Hg: El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

H1: Si se identifica los indicadores significativos de enfermedades respiratorias mediante minería de datos entonces se podrá mejorar el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco.

Queda validada y demostrada en la investigación.

H2: La aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice significativamente el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco.

Resulta validada en la investigación.

4.4 Discusión de resultados

Para la hipótesis especifica h1, el análisis realizado entre los predictores que influyen significativamente en el diagnóstico, se establece que 5 de ellos del total de 15 (vea tabla 5.1) aportan un valor de significancia

superior al 50%, esta información es proporcionada por los resultados del modelo predictivo mediante el algoritmo de árbol de decisión CyR,

En el caso de la hipótesis específica h2, la predicción que arrojan las distintas técnicas predictivas de minería de datos para el caso analizado se muestra en la tabla 5.1. Estos resultados manifiestan la existencia de evidencia suficiente para indicar que, si es posible realizar predicciones con un modelo que incluye los indicadores significativos que influyen en las enfermedades respiratorias, como es la neumonía, y que predicen con un alto grado de acierto en el diagnóstico de neumonía. Los valores alcanzados oscilaron en un valor máximo de 100% en el mejor modelo encontrado.

CONCLUSIONES

1. Los modelos de minería de datos basados en algoritmos diferentes se desarrollaron para investigar si son adecuados para describir y predecir las relaciones entre variables o indicadores predictores como por ejemplo género, edad, diagnósticos y acciones previas, así como el diagnostico de un sin número de enfermedades.
2. El modelo basado en el algoritmo Naive Bayes tuvieron un menor rendimiento para todas las pruebas que se realizaron con la información de la base de datos (historias clínicas) en comparación al modelo basado en el algoritmo de árboles de decisión con baja reducción tuvieron la mayor precisión.
3. Se considera que el modelo de árboles de decisión logró un rendimiento predictivo suficiente significativo.
4. Los patrones extraídos revelaron algunas relaciones clínicamente válidas, algunas triviales y algunas probablemente clínicamente inválidas para el diagnóstico de la neumonía como enfermedad respiratoria. Sin embargo, siempre habrá la necesidad de un profesional de la salud que tome la última decisión.
5. Finalmente, la hipótesis general de la investigación "*El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco",* se considera aceptada como consecuencia de los resultados obtenidos.

RECOMENDACIONES

1. La información obtenida de las historias clínicas (data) es de vital importancia para realizar este tipo de investigación, sin embargo la información recopilada se encontraba en formatos impresos mas no en un soporte de base de datos, lo que conllevo a un esfuerzo adicional y tiempo extra en la transferencia de la información relevante a un soporte de base de datos que permita trabajarlo, por lo que se recomienda que la organización sistematice la información histórica a fin de poder sacarle el mayor provecho posible en temas investigativos.
2. Si bien es cierto los resultados obtenidos son buenos queda un ligero sin sabor pues no todos los casos que se evaluaron de pacientes mediante el modelo de minería de datos fueron acertados, por lo que se recomienda ampliar el tamaño de la muestra para un mejor resultado predictivo en futuras investigaciones.

BIBLIOGRAFÍA

IMPRESOS

Balestrini, Miriam. (2006). Cómo se elabora el Proyecto de investigación. Caracas: Consultores Asociados.

Bernuy, A. (2018). Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la escuela profesional de ingeniería de computación y sistemas, universidad de San Martín de Porres, Lima-Perú. Investigación de pregrado, Universidad San Martin de Porras. Lima.

Candela, C. (2015). Proceso de Descubrimiento de Conocimiento para Predecir el Abandono de Tratamiento en una Entidad de Salud Pública. Lima – Perú.

Daza, A. (2016). Un modelo basado en arboles de decisión para predecir la deserción estudiantil en la educación superior privada. Universidad Cesar Vallejo.

Eapen, A. (2004). Application of data mining in medical applications. Tesis de maestría, Universidad de Waterloo. Canadá.

Eccles M, Grimshaw J, Johnston M, Steen N, Pitts N, Thomas R, Glidewell E, Maclennan G, Bonetti D and Walker A (2007). Applying psychological theories to evidence-based clinical practice: identifying factors predictive of managing upper respiratory tract infections without antibiotics. Implementation Science.

Han, J., Kamber, M. & Pei, J. (2012). Data Mining. Concepts and techniques. USA: Morgan Kaufmann. Elseiver.

Hernández, R., Fernández, C. y Baptista, M. (2010). Metodología de la Investigación. México: McGraw-Hill/Interamericana Editores.

Lambert S, Allen K, Carter R and Nolan T (2008). The cost of community managed viral respiratory illnesses in a cohort of healthy preschool-aged children. Respiratory Research 9.

MacLennan, J., Tang, Z. y Crivat, B. (2008). *Data Mining with MicrosoftSQL Server 2008*. Washington: Wiley Publishing, Inc.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72.

Turban, E., McLean, E. R. et al. (1996), Information Technology for Management: Improving Quality and Productivity, New York: John Wiley.

Simoes AF, Cherian T, Chow J, Shahid-Salles S, Laxminarayan R, John TJ (2006): Infecciones respiratorias agudas en niños. En las prioridades de control de enfermedades en los países en desarrollo. Washington: Oxford University Press.

Vela, Y. (2018). Caracterización epidemiológica de las infecciones respiratorias agudas (IRA) en hospitalización pediátrica, clínica Antioquia-Bello, Colombia, año 2016. Tesis de maestría, Universidad Nacional Autónoma de Nicaragua.

Williams B, Gouws E, Boschi-Pinto C, Bryce J and Dye C (2002). Estimates of Worldwide Distribution of Child Deaths from Acute Respiratory Infections. Lancet Infectious Diseases.

DIGITALES

1. Falkenberg Ed, Hesse W, Lind green P, Nilsson Be, Oei Jlh, Rolland C, Stamper Rk, Van Assche Fjm, Verrign-Stuart Aa And Voss K (1998). *A Framework of Information System Concepts: The FRISCO Report*. Recuperado de http://www.mathematik.unimarburg.de/~hesse/papers/fri-full.pdf

2. Organización Mundial de la Salud (2018). Infecciones Respiratorias agudas. Recuperado de https://www.who.int/es/news-room/fact-sheet/detail/the-UTRI.

ANEXO

ANEXO 1: MATRIZ DE CONSISTENCIA

TEMA: "RECONOCIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS MEDIANTE MINERÍA DE DATOS PARA MEJORAR EL DIAGNOSTICO EN PACIENTES DEL HOSPITAL DANIEL ALCIDES CARRIÓN - PASCO"

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADORES	DISEÑO METODOLÓGICO
Problema general ¿Cómo influye el reconocimiento de patrones en enfermedades respiratorias mediante minería de datos en el diagnóstico de pacientes del Hospital Daniel Alcides Carrión - Pasco? **Problemas especificas** ¿Qué variables de enfermedades respiratorias se relacionan con el diagnóstico de pacientes del Hospital Daniel Alcides Carrión - Pasco? ¿En qué medida la aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco?	**Objetivo general** Determinar la influencia de reconocer patrones en enfermedades respiratorias mediante minería de datos en el diagnóstico de pacientes del hospital Daniel Alcides Carrión – Pasco. **Objetivos específicos** Identificar las variables de enfermedades respiratorias que se relacionan con el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco. Establecer el modelo sobre enfermedades respiratorias mediante minería de datos para predecir el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco. .	**Hipótesis general** El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco. **Hipótesis especificas** Si se identifica los indicadores significativos de enfermedades respiratorias mediante minería de datos entonces se podrá mejorar el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco. La aplicación del modelo sobre enfermedades respiratorias mediante minería de datos predice significativamente el diagnostico de pacientes del del Hospital Daniel Alcides Carrión – Pasco.	**Variable Independiente** Reconocimiento de patrones en enfermedades respiratorias **Variables Dependientes** Diagnóstico de pacientes.	▪ Identificación de indicadores en enfermedades respiratorias. ▪ Modelo predictivo de enfermedades respiratorias ▪ Tipo de diagnóstico del paciente	**Tipo de Investigación** Aplicada y correlacional **Diseño de la Investigación** De tipo transeccional **Método de la Investigación** Análisis y sintesis **Población** La población a tomar en cuenta es de 648 historias clínicas de pacientes del hospital Daniel Alcides Carrión en el año 2016, periodo enero - junio. **Muestra** La muestra de estudio es de 242 historias clinicas.

HOSPITAL NACIONAL DANIEL ALCIDES CARRION

HISTORIA CLINICA

DATOS DE FILIACIÓN

Nombre: María Marcia Llangana

Edad: 24

Sexo: Femenino

Estado Civil: Casada

Ocupación: Ama de casa

Lugar de Nacimiento: Pasco

Residencia Actual: Pasco, Yanacancha

MOTIVO DE CONSULTA:

Tos y Dolor torácico

ENFERMEDAD ACTUAL

Paciente refiere como fecha real y aparente que hace 4 días aproximadamente, teniendo como causa aparente exposición al frio, presenta dolor torácico, de gran intensidad (EVA10), punzante-urente, continuo durante todo el día que se irradia a región dorsal, además presenta tos productiva durante todo el día, se acompaña con abundante expectoración de color verdoso, presencia de leve himoptismo que se acompaña de disnea de pequeños esfuerzos, presenta alza térmica, astenia y cefalea.

Refiere que el dolor sede al adoptar posición antialgica y al presionar en el lado afectado, no se relaciona con la alimentación, micción y defecación, se auto medica con paracetamol y amoxicilina por una ocasión sin presentar mejora, el cuadro se exacerba por lo cual acude a esta casa de salud. Actualmente cuadro muestra mejoría.

REVISIÓN POR SISTEMA

Respiratorio: tos productiva, hemoptisis, disnea de pequeños esfuerzos, dolor torácico.

Sistema endocrino: pérdida de peso