# UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

## FACULTAD DE INGENIERÍA

## ESCUELA DE FORMACIÓN PROFESIONAL DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

# T E S I S

# Aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023

**Para optar el título profesional de:**

**Ingeniero de Sistemas y Computación**


**Autores:**

**Bach. Carlos Daniel CHAVEZ AQUINO**

**Bach. Yerson Leoncio CRISTOBAL VICENTE**

**Asesor:**

**Msc. Hebert Carlos CASTILLO PAREDES**


**Cerro de Pasco – Perú – 2024**

# UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

## FACULTAD DE INGENIERÍA

## ESCUELA DE FORMACIÓN PROFESIONAL DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

# T E S I S

## Aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023

**Sustentada y aprobada ante los miembros del jurado:**

_____

**Mg. Melquiades Arturo TRINIDAD MALPARTIDA**
**PRESIDENTE**

_____

**Mg Lisbeth Gisela NEGRETE CARHUARICRA**
**MIEMBRO**

_____

**Mg Pit Frank ALANIA RICALDI**
**MIEMBRO**

**INFORME DE ORIGINALIDAD**

**Universidad Nacional Daniel Alcides Carrión**

**Facultad de Ingeniería Unidad de Investigación**

**INFORME DE ORIGINALIDAD N° 015-2024-UNDAC/UIFI**

La Unidad de Investigación de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión en mérito al artículo 23° del Reglamento General de Grados Académicos y Títulos Profesionales aprobado en Consejo Universitario del 21 de abril del 2022, La Tesis ha sido evaluado por el software antiplagio Turnitin Similarity, que a continuación se detalla:

Tesis:

**Aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023**

Apellidos y nombres de los tesistas:

**Bach. CHÁVEZ AQUINO, Carlos Daniel**

**Bach. CRISTOBAL VICENTE, Yerson Leoncio**

Apellidos y nombres del Asesor:

**Mg. CASTILLO PAREDES, Hebert Carlos**

Escuela de Formación Profesional

**Ingeniería Sistemas y Computación**

Índice de Similitud

**19%**

**APROBADO**

Se informa el Reporte de evaluación del software similitud para los fines pertinentes:


Cerro de Pasco, 17 de enero del 202

## DEDICATORIA

Dedicamos este trabajo primero a Dios, quien nos ha dado vida y salud para llegar a este importante momento de la formación profesional. Gracias a nuestros padres, ellos son nuestro apoyo más importante, brindándonos siempre amor y apoyo incondicional y formando nuestros buenos sentimientos, hábitos y valores. Gracias a los hermanos que han estado con nosotros, apoyado y asesorado en la búsqueda de una educación profesional. Gracias a nuestros amigos y al equipo que formamos, llegaron al final del camino y siguen siendo buenos amigos hasta el día de hoy. Gracias, nuestros maestros, por su tiempo, apoyo y paciencia, y por la sabiduría que nos han brindado a medida que desarrollamos nuestra educación profesional. Gracias a la Universidad Nacional Daniel Alcides Carrión y en especial a la Escuela de Ingeniería de Sistemas y Computación por recibirnos en sus aulas y brindarnos la oportunidad de ser parte de una generación exitosa.

**AGRADECIMIENTO**

Agradecer a Dios por protegernos y brindarnos la fuerza para terminar una meta más.

También a nuestros padres que con esfuerzo y dedicación lograron guiar nuestros pasos para poder cerrar de forma exitosa esta etapa.

A nuestros hermanos quienes siempre nos apoyaron para salir delante de los obstáculos en la vida.

A nuestros amigos con quienes compartimos experiencias y consejos.

A los Ingenieros y docentes por su paciencia y sabiduría, quienes supieron inculcarnos los conocimientos para el desarrollo de este trabajo.

Agradecemos también al Ing. Herbert Castillo por su guía y asesoramiento para la culminación de este proyecto.

Agradezco a todos que fueron apoyo para realizar este proyecto.

## RESUMEN

El trabajo de investigación que realice se titula: “Aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023” El objetivo principal fue aplicar minería de datos para mejorar la detección de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco en el año 2023. El estudio utilizado fue una prueba experimental preliminar. La población estuvo compuesta por 56 pacientes y la muestra incluyó a 20 pacientes mediante el método de muestreo por conveniencia. En la muestra no probabilística, se seleccionaron como criterios de inclusión sujetos que padecían nasofaringitis aguda (resfriado), faringitis aguda, bronquitis aguda no especificada y enfermedades no especificadas. Según los criterios de exclusión, se rechazaron los pacientes que no cumplían con la historia clínica o los criterios ambulatorios del departamento. El formulario de inscripción sirve como ayuda. Este logro se logró mejorando la detección de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco mediante la implementación de una aplicación de minería de datos que reduce efectivamente el tiempo promedio requerido para detectar enfermedades respiratorias.



**Palabras Clave.** Minería de datos, enfermedades respiratorias.

**ABSTRACT**

The research work I carried out is titled: “Application of data mining to improve the identification of respiratory diseases in the Daniel Alcides Carrión Hospital Pasco, 2023” The main objective was to apply data mining to improve the detection of respiratory diseases in the Hospital Daniel Alcides Carrión Pasco in the year 2023. The study used was a preliminary experimental test. The population was composed of 56 patients and the sample included 20 patients using the convenience sampling method. In the non-probabilistic sample, subjects suffering from acute nasopharyngitis (cold), acute pharyngitis, unspecified acute bronchitis, and unspecified diseases were selected as inclusion criteria. Based on the exclusion criteria, patients who did not meet the department's clinical history or outpatient criteria were rejected. The registration form serves as help. This achievement was achieved by improving the detection of respiratory diseases at the Daniel Alcides Carrión Pasco Hospital through the implementation of a data mining application that effectively reduces the average time required to detect respiratory diseases.

**Keywords.** Data mining, respiratory diseases.

## INTRODUCCIÓN

En Perú, Diario El Peruano (2020) demostró que el cambio climático (temperaturas más bajas) provocó un aumento de enfermedades respiratorias, siendo la más común la influenza causada por COVID-19. Los síntomas como fiebre alta, tos seca y dificultad para respirar son similares, lo que hace que el diagnóstico del problema sea más preciso.

Por otro lado, en una publicación realizada por el Diario Peru21 (2019) mencionó alrededor de los meses de abril y agosto los meses de abril y agosto (épocas de temperaturas bajas), los pacientes de distintas edades presentaron 42,073 casos de Infecciones Respiratorias Agudas (IRA), lo cual fue un factor determinante en el aumento de casos de dicho grupo de enfermedades:

Capítulo I: Incluye los siguientes apartados: Identificación y determinación del problema, delimitación de investigación, formulación del problema, formulación de objetivos, justificaciones de la investigación y limitaciones de la investigación".

Capítulo II: Fundamentos teóricos y científicos, Definir términos, identificar hipótesis y variables, y finalmente definiciones operativas.

Capítulo III: Tipos de investigación, métodos de investigación, diseño de investigación, conjunto principal y muestra, métodos y herramientas de recolección de datos, técnicas de procesamiento y análisis de datos, procesamiento de datos estadísticos, selección de herramientas de investigación, validación y confiabilidad, ética y pautas de investigación.

Capítulo IV: Resultados y discusión incluye las siguientes partes: descripción del trabajo, análisis e interpretación de resultados, prueba de hipótesis y discusión de resultados. Finalmente, proporcionamos conclusiones, recomendaciones, referencias y anexos.

**El autor.**

**ÍNDICE**

DEDICATORIA.

AGRADECIMIENTO

RESUMEN

ABSTRACT

INTRODUCCIÓN

ÍNDICE

CAPITULO I

PROBLEMA DE INVESTIGACIÓN

1.1. Identificación y determinación del problema ..... 1

1.2. Delimitación de la investigación. ..... 3

1.3. Formulación del problema ..... 3

1.3.1. Problema general ..... *3*

1.3.2. Problemas específicos. ..... *3*

1.4. Formulación de objetivos ..... 4

1.4.1. Objetivo General. ..... *4*

1.4.2. Objetivos específicos. ..... *4*

1.5. Justificación de la investigación. ..... 4

1.6. Limitaciones de la investigación ..... 4

CAPITULO II

MARCO TEÓRICO

2.1. Antecedentes de estudio. ..... 6

2.2. Bases teóricas – científicas. ..... 12

2.3. Definición de términos básicos. ..... 23
2.4. Formulación de Hipótesis ..... 24
2.4.1. Hipótesis General ..... *24*
2.4.2. Hipótesis Específicas ..... *24*
2.5. Identificación de Variables ..... 25
2.6. Definición Operacional de variables e indicadores ..... 25

CAPITULO III

METODOLOGÍA Y TECNICAS DE INVESTIGACIÓN

3.1. Tipo de investigación ..... 26
3.2. Nivel de investigación ..... 26
3.3. Métodos de investigación ..... 26
3.4. Diseño de investigación ..... 27
3.5. Población y muestra ..... 27
3.6. Técnicas e instrumentos de recolección de datos. ..... 27
3.7. Selección, validación y confiabilidad de los instrumentos de investigación. ..... 28
3.8. Técnicas de procesamiento y análisis de datos. ..... 29
3.9. Tratamiento Estadístico. ..... 31
3.10. Orientación ética filosófica y epistémica. ..... 32

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1. Descripción del trabajo de campo ..... 33
4.2. Presentación, análisis e interpretación de resultados ..... 34
4.3. Prueba de Hipótesis ..... 37
4.4. Discusión de resultados ..... 40

CONCLUSIONES

RECOMENDACIONES

REFERENCIAS BIBLIOGRÁFICAS

ANEXOS

## ÍNDICE DE TABLAS

Tabla 1Definición Operacional de Variables..... 25

Tabla 2.Tabla de validación ..... 29

Tabla 3.Medidas descriptivas del indicador..... 35

Tabla 4.Medidas descriptivas de la dimensión 2 ..... 36

Tabla 5.Prueba de normalidad de la Dimensión 1..... 37

Tabla 6.Prueba de rangos con signo de Wilcoxon de la Dimensión 1 ..... 38

Tabla 7.Prueba Z de la Dimensión 1..... 38

Tabla 8.Prueba de normalidad de la Dimensión 2..... 39

Tabla 9.Prueba de rangos con signo de Wilcoxon de la Dimensión 2 ..... 39

Tabla 10.Prueba de normalidad de la Dimensión 2..... 40

## ÍNDICE DE FIGURAS

Figura 1Proceso de descubrimiento de conocimiento. ..... 12

Figura 2Algoritmo Naive Bayes basado en las probabilidades ..... 17

Figura 3.Ejemplo de Árbol de decisión. ..... 18

Figura 4.Representación de red neuronal y de regresión logística. ..... 19

Figura 5Ubicación. ..... 34

Figura 6.Nivel de morbilidad - Antes y después de la implementación ..... 35

Figura 7.Dimensión de tiempo promedio de diagnostico ..... 36

# CAPITULO I

# PROBLEMA DE INVESTIGACIÓN

## 1.1. Identificación y determinación del problema

Para 2020, el progreso ha aumentado en el proceso de enfermedad respiratoria diagnóstica, como mejorar su tecnología, como la imagen y el examen biológico, pero la integridad del paciente aún daña estas dificultades. Del mismo modo, el Instituto de Investigación Médica de Texas encontró que la displasia causó 400,000 muertes, lo que resultó en errores médicos preventivos.

"Además, la Organización Mundial de la Salud, también conocida como OMS, informó que más de 138 millones de pacientes sufrieron errores médicos en 2019, lo que sugiere que este sigue siendo el caso". Las presentes estadísticas fueron referentes países de bajo y medio nivel de ingresos, abarcando el 80% de los habitantes globalmente (Gestión, 2019). De igual manera, en Estados Unidos, dicha organización llevo a cabo un estudio donde se evidenciaron errores de diagnóstico, generando alrededor de un 10% de mortalidad (Índice de personas que mueren) en los hospitalizados, lo cual represento entre el 6 % y 17 % de daños a la integridad del paciente (OMS, 2019).

En Cuba, en el Centro Provincial de Información de Ciencias Médicas de Camagüey, se realizó un estudio logrando detectar que las Infecciones

Respiratorias Agudas (IRA), son causa principal de morbilidad (Índice de personas que se enferman); además fueron las más consultadas en personas de todas las edades, donde se determinó que estas infecciones afectaban principalmente a menores de 15 años (Reus y Ortiz, 2013).

En Perú, Diario El Peruano (2020) demostró que el cambio climático (temperaturas más bajas) provocó un aumento de enfermedades respiratorias, siendo la más común la influenza causada por COVID-19. Los síntomas como fiebre alta, tos seca y dificultad para respirar son similares, lo que hace que el diagnóstico del problema sea más preciso.

Además, El Ministerio de Salud (MINSA) difundió que, en el año 2019, la mortalidad de la neumonía ha incrementado, alcanzado a un total de 38 mil casos nuevos en los niños menores de 5 años, siendo la población con nivel de riesgo alto (Expreso, 2019).

Por otro lado, en una publicación realizada por el Diario Peru21 (2019) mencionó alrededor de los meses de abril y agosto los meses de abril y agosto (épocas de temperaturas bajas), los niños de esas mismas edades, presentaron 42,073 casos de Infecciones Respiratorias Agudas (IRA), lo cual fue un factor determinante en el aumento de casos de dicho grupo de enfermedades.

Esta situación también ocurre en el Hospital Daniel Alcides Carrión de Pasco, donde “se procesa un sin fin de información y al final se registran y almacenan decenas de facturas, préstamos, tratamientos y servicios. Toda esta información puede ayudarte. Informarse. Por ejemplo, las organizaciones pueden saber qué enfermedades son más comunes por región o grupo de edad. Para hacer esto, necesita herramientas para administrar todos sus datos y convertirlos en información a través de modelos y búsquedas heurísticas. Todavía queda mucho trabajo por hacer, incluso si se puede utilizar para identificar y regular patrones de síntomas y enfermedades, y si se puede utilizar para crear sistemas de información que respalden los procesos de atención

médica. Por tanto, el propósito de este estudio es demostrar cuántos modelos de apoyo que reflejan el comportamiento del paciente pueden servir como base para redirigir recursos".

## 1.2. Delimitación de la investigación.

### 1.2.1. Delimitación espacial

Estaré realizando una investigación utilizando datos del 2023 sobre pacientes con enfermedades respiratorias en el Hospital Daniel Alcides Carrión en Pasco.

### 1.2.2. Delimitación temporal

Investigación de interpretación de información realizada como parte del proceso de investigación de recopilación de datos de 2022 a 2023.

### 1.2.3. Delimitación conceptual

Se buscan conceptos para minería de datos y detección de enfermedades respiratorias.

## 1.3. Formulación del problema

### 1.3.1. Problema general

¿Se podrá aplicar la minería de datos para la mejora en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023?

### 1.3.2. Problemas específicos

¿Se podrá disminuir el nivel de morbilidad aplicando la minería de datos en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023?

¿Se podrá disminuir el tiempo promedio aplicando la minería de datos para identificar la existencia de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023?

## 1.4. Formulación de objetivos

### 1.4.1. Objetivo General

Aplicar la minería de datos para la mejora en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

### 1.4.2. Objetivos específicos

Disminuir el nivel de morbilidad aplicando la minería de datos en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

Disminuir el tiempo promedio aplicando la minería de datos para identificar la existencia de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

## 1.5. Justificación de la investigación

Hoy en día, las organizaciones son juzgadas no sólo por la calidad de sus productos y servicios, sino también por lo bien que se comunican con los clientes, empleados y socios. Sin embargo, la mayoría de las organizaciones cuentan con una gran cantidad de información desconocida y no utilizada. Por esta razón se desarrolló la minería de datos. Busca no sólo extraer información de los datos, sino también extraer información real que pueda proporcionar una ventaja competitiva en la toma de decisiones organizacionales.

Actualmente, la situación de las enfermedades respiratorias en el Hospital Daniel Alcides Carrión de Pasco ha cambiado dramáticamente, causando confusión en los diagnósticos de los médicos y que los informes no se reciben a tiempo para tomar decisiones. Por lo tanto, el propósito de este estudio y el modelo resultante es ayudar a los pacientes hospitalizados a realizar diagnósticos oportunos y precisos y a tomar decisiones médicas.

## 1.6. Limitaciones de la investigación

Las limitaciones en la realización de este estudio están relacionadas con el escepticismo de los responsables de compartir datos y bases de datos de

pacientes respiratorios, porque se sabe que se trata de información confidencial, por lo que se decidió anonimizar los datos. para cada paciente. También existen limitaciones financieras, ya que todo el programa de investigación se autofinancia.

# CAPITULO II

# MARCO TEÓRICO

## 2.1. Antecedentes de estudio.

### 2.1.1. A nivel Internacional

Según Oviedo, Oviedo y Vélez (2023) en su trabajo de investigación titulado “Minería de Datos: Aportes y tendencias en el Servicio de Salud de Ciudades Inteligentes” en su trabajo de investigación titulado “Minería de Datos: Aportes y tendencias en el Servicio de Salud de Ciudades Inteligentes” Entre las muchas aplicaciones de la minería de datos, su contribución a la atención sanitaria en las ciudades inteligentes es particularmente importante. Estas aplicaciones tienen como objetivo mejorar la calidad de vida de la población, prevenir enfermedades, facilitar la toma de decisiones y analizar datos de los centros sanitarios. Para apoyar el desarrollo de ciudades inteligentes, este artículo examina los avances y tendencias en la minería de datos médicos. Los mayores avances en la minería de datos incluyen una variedad de tecnologías, técnicas y plataformas que ya se utilizan en la industria médica. Según estas tendencias, se pueden identificar varios problemas, entre ellos: B. Herramientas que soportan el análisis de textos e imágenes, métodos de indexación y procesamiento de datos no estructurados y minería multimedia.

Según Rojas y Sebastián (2017) en su trabajo de investigación titulado “Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia” El proyecto se basa en una aplicación de minería de datos utilizando el algoritmo de clustering K- Means, y a través del análisis de datos se puede construir un modelo descriptivo para identificar posibles patrones de comportamiento de enfermedades respiratorias en la ciudad de Bogotá. El grupo creado con la herramienta RapidMiner es un conjunto de datos de cinco años, de 2012 a 2016, que incluye 184 casos de diagnóstico de enfermedades respiratorias, correspondientes a pacientes de 0 a 5 años, según la integridad y la correlación de variables. . La fuente de información para la selección de género es el SISPRO kutio, un sistema integrado de información sobre el seguro social. Además, a medida que la información producida por los sistemas y plataformas de salud crece exponencialmente en el tiempo, existe la necesidad de analizar la información mediante clusters para que estos datos acumulados puedan luego transformarse en información y convertirse en insumos para la toma de decisiones importantes. Diagnóstico de enfermedades respiratorias. A continuación, se presenta una metodología temática que consta de pasos que permiten la implementación de aplicaciones de minería de datos. En primer lugar, implica obtener la base de datos y sus variables, en segundo lugar, elegir herramientas de análisis de datos, en tercer lugar, utilizar la base de datos en herramientas analíticas y luego aplicar algoritmos de agrupamiento para identificar patrones de comportamiento. Obtenga los resultados utilizando los patrones identificados y finalmente escriba los resultados que obtenga. Finalmente, cabe aclarar que las variables tomadas durante el desarrollo fueron datos no sensibles y no se obtuvieron de diferentes fuentes de datos para mantener la seguridad e integridad del paciente. Asimismo, se ve la necesidad de realizar análisis de la información mediante clustering, debido al crecimiento

exponencial de datos que generan los sistemas y plataformas del sector salud con el paso del tiempo, permitiendo que estos datos organizados en los clústeres finalmente se conviertan en información y sea un insumo para la toma de decisiones enfocado a los diagnósticos de las enfermedades respiratorias. A continuación, se relaciona la metodología que consiste una serie de pasos que permiten lograr la aplicación de minería de datos. Primero, consiste en la obtención de bases de datos y sus variables, segundo la selección de la herramienta de análisis de datos, tercero la aplicación de las bases de datos en la herramienta de análisis, luego la aplicación algoritmos de clustering que permitan identificar los patrones de comportamiento, seguido de la obtención de resultados mediante los patrones ya identificados y por último realizar la documentación sobre los resultados obtenidos. Finalmente, cabe aclarar que las variables que se contemplan para el desarrollo, son datos que no son sensibles y que tampoco son suministradas por las diferentes fuentes de información con el fin de mantener la seguridad e integridad de los pacientes.

Según Bautista (2010) en su trabajo de investigación titulado “Uso de minería de datos en la detección temprana y prevención de complicaciones de enfermedades en el sistema de salud Colombiano” tuvo como objetivo pretendió Incluye complicaciones relacionadas con enfermedades respiratorias identificadas por las autoridades sanitarias colombianas y detectadas mediante minería de datos. Inicialmente, estas dificultades eran médicas y administrativas. En este proyecto, buscamos descubrir la probabilidad de que los pacientes tuvieran una enfermedad respiratoria preexistente. Además, se buscaron asociaciones entre las complicaciones y los departamentos que tratan a los pacientes. Para resolver el problema propuesto originalmente, se propone utilizar reglas de agrupación y asociación como métodos de minería de datos. Se decidió introducir un modelo de conglomerados para conocer la relación entre la frecuencia de uso de los servicios médicos y la aparición de

complicaciones. En base a esto, se desarrolló un proceso de extracción para utilizar estos valores y algunas de las características utilizadas en los grupos de características como contexto para determinar qué tipos de complicaciones causaron las complicaciones. Después del estudio, los resultados mostraron una correlación entre las complicaciones de los pacientes y los datos clínicos, y una correlación entre la demografía de los pacientes y la probabilidad de complicaciones.

Según Jiménez (2019) en su trabajo de investigación titulado “Aplicación de analítica de datos para predicción de infección respiratoria aguda en Colombia” como objetivo principal fue implementar un modelo predictivo de infección respiratoria aguda en Colombia acorde con las fuentes de datos pertinentes y disponibles para el proyecto de 2013 al 2018. Este proyecto propuso una solución al problema del pronóstico de Infección Respiratoria Aguda (IRA) en Colombia a través del diseño de múltiples modelos por departamento, superando el alcance establecido inicialmente de un solo modelo. Estos modelos permitieron evaluar y reconocer bajo cuáles condiciones técnicas es posible hacer pronósticos más acertados. La principal fuente de datos que puede ser utilizada para la predicción de IRA en Colombia debe ser la variable histórica de la enfermedad recopilada por el Instituto Nacional, pues en varios casos en los que fue utilizada únicamente las métricas de los errores disminuyeron. No está descartado el uso de variables externas a la enfermedad, ya que en casos multivariados también se obtuvieron pronósticos acertados, solo que dependió de la ubicación geográfica o departamento. Además, se construyó un visualizador de datos, que representa una herramienta crucial en la interpretación de resultados. Su diseño se realizó teniendo en cuenta la importancia de utilizar interfaces no solo amigables sino también adecuadas para la correcta interpretación de los modelos. Por lo tanto, el visualizador

constituye el intermediario entre el modelo y el usuario que puede ser un profesional de salud, un experto, un analista de datos o un ciudadano.

### 2.1.2. A nivel Nacional

Según Alva y Cruz (2021) en su trabajo de investigación titulado "Aplicación de minería de datos para mejorar el diagnóstico de un grupo de enfermedades respiratorias en un hospital de Trujillo" El objetivo general es utilizar aplicaciones de minería de datos para mejorar el diagnóstico de enfermedades respiratorias del grupo CAP III en la atención primaria de salud en la capital Trujillo. En este caso se realizó un estudio preexperimental. Se utilizaron herramientas de recolección de datos como cuadernos y discos de observación, y la confiabilidad se garantizó mediante juicio de expertos utilizando el coeficiente V de Aiken, que tiene una validez del 95%. El proyecto utilizó Pentaho Data Integration, Pentaho Server, MySQL y otras tecnologías. Una vez desarrollado el software, siga los pasos de la metodología CRISP-DM, incluida la comprensión empresarial, la comprensión de datos, la preparación de datos, el modelado, la evaluación y la implementación. Hubo un total de 56 pacientes, de los cuales 19 fueron examinados. Los resultados obtenidos luego de implementar la aplicación de minería de datos fueron que la morbilidad aumentó en un 4.96% y el tiempo promedio requerido para detectar enfermedades respiratorias disminuyó en 20 minutos. Además, el costo promedio disminuyó en S/327.95. Actividades de diagnóstico en este grupo. Finalmente, el Centro Primario Metropolitano CAP III Trujillo ha logrado importantes avances en el diagnóstico de un grupo de enfermedades respiratorias.

Según Jara (2023) en su trabajo de investigación titulado "Identificación automática de neumonía mediante el procesamiento digital del sonido" La neumonía es ahora la causa más común de muerte en la niñez y mata a 1,1 millones de niños menores de 5 años cada año. Para un diagnóstico rápido de

esta enfermedad, se puede escuchar los pulmones con un estetoscopio, lo que permite sentir los ruidos respiratorios e identificar signos patológicos. El objetivo de este trabajo es detectar automáticamente la neumonía mediante procesamiento de audio digital. El estudio primero desarrolló un protocolo de recolección de sonidos respiratorios y luego los convirtió en imágenes. Para obtener un espectrograma para cada imagen, el espectrograma es una imagen procesada con Keras. También construimos una red neuronal convolucional y comenzamos a entrenar con datos de imágenes con un valor inicial de 200 épocas. Los resultados fueron satisfactorios con exactitud, precisión y sensibilidad del 75%, 69% y 75% respectivamente. Finalmente, el método fue evaluado y mostró un buen desempeño en la detección automática de neumonía.

### 2.1.3. A nivel Local

Según Curi (2020) “Reconocimiento de patrones en enfermedades respiratorias mediante minería de datos para mejorar el diagnostico en pacientes del Hospital Daniel Alcides Carrión – Pasco” El propósito era determinar "el impacto de la detección de patrones de enfermedades respiratorias mediante la extracción de datos en el diagnóstico de pacientes en el Hospital Daniel Alcides Carrión de Pasco". En el desarrollo de la investigación se utilizan métodos de análisis y síntesis por lo que se divide el trabajo 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). 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. 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

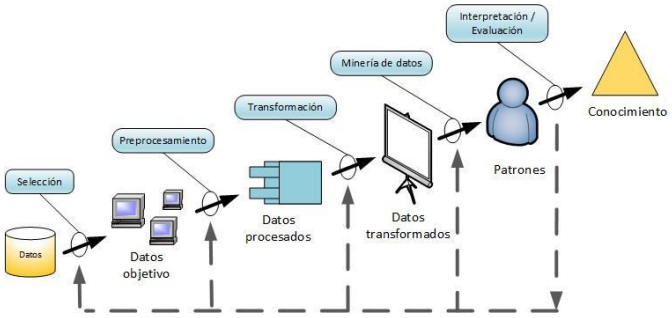
es la neumonía, y que predicen con un alto grado de acierto en el diagnóstico de neumonía. Los valores alcanzados oscilaron en un valor máximo de 100% en el mejor modelo encontrado". En conclusión, la hipótesis general de la investigación "El reconocimiento de patrones en enfermedades respiratorias mediante minería de datos mejora el diagnóstico de pacientes del Hospital Daniel Alcides Carrión – Pasco", se considera aceptada como consecuencia de los resultados obtenidos.

## 2.2. Bases teóricas – científicas.

### 2.2.1. Minería de datos.

Con el desarrollo de las tecnologías de la información se han desarrollado diversas bases de datos, acumulándose una enorme cantidad de información en diversos campos. La investigación en bases de datos y tecnología de la información ha creado enfoques para almacenar y procesar esta valiosa información para tomar decisiones mejores y más importantes. La minería de datos implica extraer información útil de big data. También conocido como proceso de recuperación de información extracción de información a partir de datos, extracción de información o análisis de datos/patrones.

**Figura 1 Proceso de descubrimiento de conocimiento.**




Una definición de minería de datos la obtenemos de Han, J. et. al. (2012). La minería de datos consiste en encontrar patrones interesantes en grandes cantidades de datos. Los procesos de minería de datos suelen incluir limpieza de datos, integración de datos, selección de datos, transformación de datos,

reconocimiento de patrones, evaluación de patrones y representación de datos. Datos para descubrir patrones ocultos en su base de datos. El patrón elegido debe ser racional de alguna manera, por ejemplo desde una perspectiva financiera o desde una perspectiva de seguridad del paciente.

Además del término "minería de datos", en la literatura se utilizan otros términos más o menos idénticos, como aprendizaje automático, análisis predictivo y minería de datos (KDD). La minería de datos transforma grandes cantidades de datos en información. Los motores de búsqueda como Google reciben cientos de millones de consultas cada día. Cada encuesta puede considerarse un evento donde los usuarios describen sus necesidades de información. ¿Qué cosas nuevas y útiles pueden aprender los motores de búsqueda con el tiempo a partir de una

¿Hay muchas solicitudes de los usuarios? Curiosamente, algunos patrones encontrados en las búsquedas de los usuarios pueden revelar información valiosa que no se puede obtener leyendo un solo dato. Por ejemplo, Google Flu Trends utiliza términos de búsqueda específicos como indicadores de brotes de influenza. Los estudios han encontrado una fuerte correlación entre la cantidad de personas que buscan información sobre la gripe y la cantidad de personas que realmente tienen síntomas de la gripe. Este ejemplo muestra cómo la minería de datos transforma grandes cantidades de datos en información que ayuda a resolver los problemas globales actuales.

Ciclo de la minería de datos

Según MacLennan, J. et al (2008) el ciclo del proyecto de minería “A minería de datos consta das seguintes etapas, en función do propósito da minería de datos”:

**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.

- **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 Algoritmo Naive Bayes basado en las probabilidades**

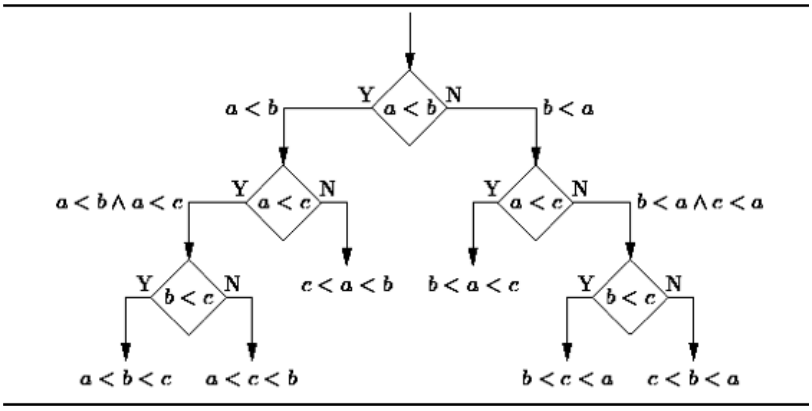
$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

- P(A): Probabilidad de A
- P(R|A): Probabilidad de que se de R dado A
- P(R): Probabilidad de R
- P(A|R): Probabilidad posterior de que se de A dado R

- **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 sobre representar 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 árbol de decisión es uno de los algoritmos más populares porque es rápido, fácil de entender y preciso si se usa correctamente.

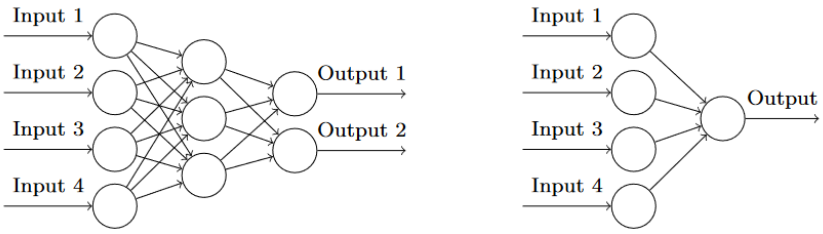
**Figura 3. Ejemplo de Árbol de decisión.**




- **Algoritmo de red neuronal:** “Los algoritmos de redes neuronales son redes neuronales artificiales que imitan la forma en que funciona la mente humana cuando se enfrenta a un problema. Se analizan todas las combinaciones de entradas y salidas y se asignan pesos a sus relaciones. También busca no sólo la entrada, sino también el conjunto de entradas asociadas con la salida. Además, hay una capa oculta de nodos entre las entradas y las salidas, por lo que las entradas no están conectadas directamente a las salidas”.
- **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, 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.

**Figura 4. Representación de red neuronal y de regresión logística.**




a) Red neural b) 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.

### 2.1.1. Enfermedades respiratorias

"El tracto respiratorio superior incluye la nariz, la boca, las fosas nasales, la faringe y la laringe, y el tracto respiratorio inferior incluye la tráquea, los bronquios principales y los pulmones".

Esta estructura dirige el aliento de aire exterior hacia los pulmones, donde se produce 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). Se pueden nombrar específicamente las infecciones en determinadas zonas del tracto respiratorio superior. Ejemplos de estos pueden ser rinitis (inflamación de las fosas nasales), sinusitis (inflamación de los senos nasales o rinosinusitis):

- "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.

**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
- 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.

## 2.3. Definición de términos básicos.

- **Árboles 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.

- **Clasificación**

  Forma de análisis de datos que permiten extraer modelos que describen las clases importantes de los datos.

- **Enfermedades:**

  La OMS define enfermedad como "Alteración o desviación del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible"

- **Identificación**

  Identificación es la acción y efecto de identificar o identificarse (reconocer si una persona o una cosa es la misma que se busca, hacer que dos o más cosas distintas se consideren como una misma, llegar a tener las mismas creencias o propósitos que otra persona, dar los datos necesarios para ser reconocido).

- **Morbilidad:**

  Se refiere a la presentación de una enfermedad o síntoma de una enfermedad, o a la proporción de enfermedad en una población. La morbilidad también se refiere a los problemas médicos que produce un tratamiento.

- **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

## 2.4. Formulación de Hipótesis

### 2.4.1. Hipótesis General

La aplicación de minería de datos mejora la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

### 2.4.2. Hipótesis Específicas

La aplicación de minería de datos disminuirá el nivel de morbilidad de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

La aplicación de minería de datos disminuirá el tiempo promedio de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

## 2.5. Identificación de Variables

### 2.5.1. Variables independientes

Aplicación de minería de datos.

### 2.5.2. Variables dependientes

Identificación de enfermedades respiratorias.

## 2.6. Definición Operacional de variables e indicadores

**Tabla 1 Definición Operacional de Variables**

| VARIABLE | DIMENSIONES |
|---|---|
| INDEPENDIENTE<br>Aplicación de minería de datos | - Análisis. |
| DEPENDIENTE<br>Identificación de enfermedades respiratorias | - Nivel de morbilidad<br>- Tiempo promedio de diagnóstico |

# CAPITULO III

# METODOLOGÍA Y TECNICAS DE INVESTIGACIÓN

## 3.1. Tipo de investigación

El tipo de la investigación del presente estudio es aplicada. Según Ortega (2021), es el estudio científico que busca resolver un problema o planteamiento específico, que se caracteriza por buscar la aplicación o utilización de los conocimientos que se adquieren.

## 3.2. Nivel de investigación

El método analítico es aquel método de investigación que consiste en la desmembración de un todo descomponiéndolo en sus partes o elementos para observar las causas, naturaleza y los efectos.(Hernández Sampieri & Mendoza Torres, 2018)

## 3.3. Métodos de investigación

La investigación que realizo es de integración analítica, que busca encontrar la causa de un evento estableciendo relaciones causales. En este sentido, la investigación descriptiva puede lograr ambas cosas de la determinación de las causas (investigación post facto), como de los efectos (investigación experimental), mediante la prueba de hipótesis. Sus resultados y conclusiones constituyen el nivel más profundo de conocimientos" según(Hernández Sampieri & Mendoza Torres, 2018).

## 3.4. Diseño de investigación

Experimental de tipo preexperimental. (Hernández Sampieri & Mendoza Torres, 2018)


Dónde:

$O0$: identificación de enfermedades respiratorias antes de la implementación de una aplicación de minería de datos.

$X$: aplicación de minería de datos.

$O1$: identificación de enfermedades respiratorias después de la implementación de una aplicación de minería de datos.

## 3.5. Población y muestra

### 3.5.1. Población

Según Hernández Sampieri (2018) menciona que: “es un conjunto de individuos que se hallan en un determinado sector y que nos apoya para adquirir la muestra y los resultados”.

La investigación abordada presenta a 56 pacientes.

### 3.5.2. Muestra

La muestra está formada por 20 pacientes, y se utiliza como criterio de inclusión una técnica de “muestreo por conveniencia” para una muestra no probabilística. Nasofaringitis aguda (resfriado), faringitis aguda, no especificada y Bronquitis aguda, no especificada, que tienen un historial médico en la entidad y hayan sido atendidos por consulta externa, en lo correspondiente a criterios de exclusión se descartaron aquellos pacientes que no cumple los criterios.

## 3.6. Técnicas e instrumentos de recolección de datos.

De acuerdo con Huaman (2005) en su libro “Manual de técnicas de investigación Conceptos y Aplicaciones” definió que el fichaje es una técnica auxiliar que consisten en la recolección o recopilación de la información

relevante de un estudio, volviéndolo un instrumento valioso, ahorra dinero y tiempo, la elaboración se realiza por ficha o formatos produciendo un valor propio, hoy en día es común recolectar dicha información útil para la investigación desde una base de datos, aunque las fichas de forma tradicional son cartulinas. Además, según Taffarel (2009) en su libro “La creación del conocimiento” indicó que el fichaje consiste en extraer y recopilar de manera ordenada la información que se requiere saber, permitiendo recolectar los datos para los indicadores de la investigación mencionados anteriormente.

Según Carrasco (2005) en su libro “Metodología de la investigación científica” definió que la ficha de observación registra los datos que se obtienen del contacto directo entre el observador y la realidad observada. Se elaboro una ficha de observación para el indicador de tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias.

### 3.7. Selección, validación y confiabilidad de los instrumentos de investigación.

Para confirmar la validez de los instrumentos, se utilizó la técnica del juicio de expertos, que según Escurra (1988) en su artículo titulado “Cuantificación de la validez de contenido por criterio de jueces” consiste en requerir la aceptación o rechazo de los instrumentos por parte de varios expertos en el tema, cuyo número puede variar dependiendo de cada investigación y para la confiabilidad se utilizó el coeficiente V de Aiken el cual computa a partir de un dato conseguido sobre la suma máxima de los valores posibles, los cuales pueden ser calculados utilizando las valoraciones de un grupo de expertos con relación a los ítems, dichas valoraciones pueden ser dicotómicas donde reciben valores de (0 o 1) o politómicas con valores de (0 a 5).

**Tabla 2. Tabla de validación**

| Experto | Ficha de registro |
|---|---|
| Experto 1 | 92 |
| Experto 2 | 100 |
| Experto 3 | 82 |
| Total | 91.33 |

En la tabla 2 visualizada se aprecia la ficha de registro que fue validada por los tres expertos y la calificación obtenida de la evaluación del indicador mencionado, obteniendo un promedio de 90.47% de validez, la cual se encuentra en la sección de anexos de la presente investigación.

### 3.8. Técnicas de procesamiento y análisis de datos.

El análisis se realizó mediante una entrevista no estructurada al director del hospital, señor Daniel Alcides Carrión Pasco, para establecer su visión, situación actual y problemática real, a partir del panorama general. Se realiza un análisis descriptivo en respuesta a presentaciones anteriores. Sobre el tratamiento. Aquí se recopila información sobre el proceso desde una perspectiva cuantitativa, como por ejemplo: Utilizar cuestionarios y formularios de registro para mejorar los diagnósticos escritos, como el nivel de incidencia mensual de enfermedades y el tiempo promedio requerido para su identificación, la presencia o ausencia mensual de enfermedades aeróbicas y el costo promedio mensual de las enfermedades respiratorias. herramienta. Se utiliza un método CRISP-DM de cinco pasos para implementar la aplicación de minería de datos. Esta es la fase de inteligencia empresarial e incluye recopilar información sobre el hospital, establecer objetivos comerciales, establecer estándares de desempeño y evaluar la salud del hospital. Identificar requisitos, limitaciones, riesgos y oportunidades. Además, se analizan los términos de negocio (esto es importante ya que ayuda a mantener las relaciones dentro de

la organización), se crea un análisis de costo-beneficio y se determinan la estrategia de minería de datos y los objetivos a utilizar. Plan de negocios de equipos mineros. Al final de la sesión se creará un plan de proyecto de aplicación para el desarrollo de este semestre, enfocado en la evaluación de equipos y tecnologías.

El segundo paso es aplicar la información de los datos. Recopilación, descripción de los datos adquiridos, análisis de datos y control de calidad de los datos finales.

El tercer paso es preparar los datos seleccionando la eliminación e inclusión de datos, la limpieza de salida, el origen del objeto, la inclusión y generación del formato de datos.

El cuarto paso es el modelado de materiales. Se definen los métodos utilizados para el modelado, se realizan configuraciones de prueba y, como resultado de estas actividades, se desarrollan modelos de materiales y configuraciones de parámetros, la conclusión también está escrita. Estos son los datos de muestra creados en este paso.

En la quinta fase se realizará una evaluación para definir los criterios de desempeño y elaborar un resumen de las acciones, decisiones y recomendaciones realizadas en cada fase del proyecto. Los pasos finales incluyen la creación del modelo de datos, la preparación del plan de instalación más exacto y preciso, la preparación del modelo operativo y el plan de mantenimiento, la preparación del informe final para presentar el proyecto y la evaluación final cuando sea hecho. Del programa. Tras el despliegue del modelo

Se realiza una evaluación previa al despliegue para obtener la información necesaria y utilizar las herramientas de evaluación post- despliegue de los indicadores mencionados anteriormente.

## 3.9. Tratamiento Estadístico.

### Análisis Descriptivo

En esta investigación se implementará una aplicación de minería de datos para evaluar el nivel de morbilidad, tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias y el costo promedio de diagnóstico de un grupo de enfermedades respiratorias en los pacientes del Hospital Daniel Alcides Carrión Pasco. Para realizar la medición de los indicadores propuestos se realizaron instrumentos como ficha de observación y de registro el cual se aplicará una evaluación antes de la implementación que permitirá conocer los objetivos mencionados anteriormente. Posteriormente se implementará una aplicación de minería de datos durante un mes, este tiempo permitirá que la aplicación influya en los indicadores anteriormente mencionados. Luego de la implementación se realizará la evaluación después de la implementación, para registrar la variación en los indicadores. Estos resultados se verán representados mediante gráficos de barras, gráficos lineales o tablas para cada de indicador, para su correcto análisis y evaluación.

### Análisis inferencial

Los datos obtenidos en esta investigación se realizarán de la siguiente manera: en primer lugar, se optará por usar la técnica de prueba de normalidad para saber si los recopilado posee o no una distribución normal, esta misma será realizada con la prueba Shapiro Wilk utilizada para muestras menores a 50 como lo menciona Romero (2016), contamos con una población de 40 y se utilizará la herramienta SPSS v26, esta prueba se realizará para cada uno de los indicadores.

Luego se determinará si lo recopilado posee o no una distribución normal, de acuerdo en ello se procederá a realizar la prueba de hipótesis correspondiente por cada indicador, usando la técnica T de Student o

Wilconxon, dependiendo del resultado obtenido de la prueba de normalidad, se determinará que hipótesis es la que se acepta.

### 3.10. Orientación ética filosófica y epistémica.

Este trabajo se lleva a cabo de acuerdo con los estándares establecidos en directrices tales como planes y programas de investigación y desarrollo, y de acuerdo con las consideraciones éticas y los principios éticos de la investigación.

# CAPITULO IV

# RESULTADOS Y DISCUSIÓN

## 4.1. Descripción del trabajo de campo

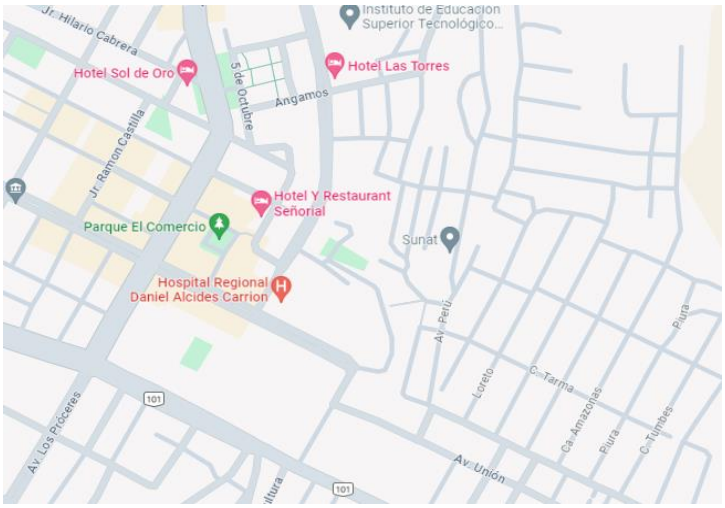
**Misión del Hospital Daniel Alcides Carrión:**

La misión del hospital es ser un hospital de referencia regional dependiente del Ministerio de Salud que brinde servicios médicos integrales y profesionales a los residentes.

**Visión del Hospital Daniel Alcides Carrión:**

Como hospital líder en la región, que brinda un servicio médico integral profesional, los recursos humanos son considerados la parte más valiosa de su organización y brindan servicios médicos de calidad a las personas.

**Figura 5 Ubicación.**




## 4.2. Presentación, análisis e interpretación de resultados

### 4.2.1. Análisis Descriptivo.

En la presente investigación se implementó una aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco para el cual se aplicó una evaluación antes de la implementación, en donde se ejecutaron los indicadores planteados que permitieron conocer el diagnóstico de un grupo de enfermedades respiratorias, se procedió con la implementación de la aplicación de minería de datos y después de la implementación se realizó nuevamente la evaluación. El resultado que se obtuvo al procesar la información recolectada se puede encontrar en la sección de anexos de este informe.

**Dimensión 1: Nivel de morbilidad**

**Análisis Descriptivo**

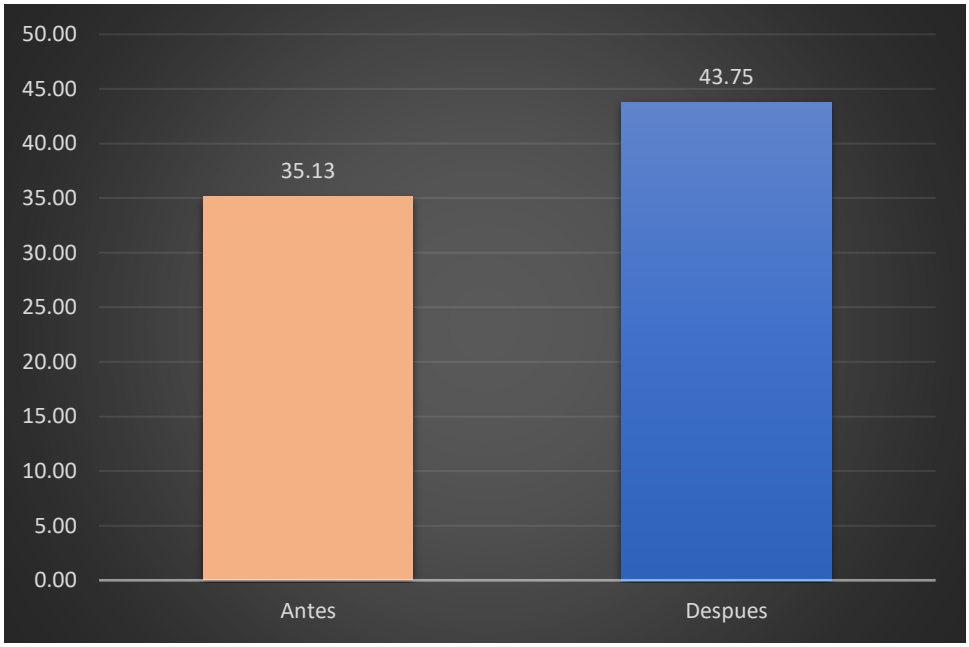
**Tabla 3. Medidas descriptivas del indicador**

**Estadísticos descriptivos**

| | N | Mínimo | Máximo | Media |
|---|---|---|---|---|
| Antes | 20 | 7,69 | 61,00 | 35,1345 |
| Despues | 20 | 9,09 | 100,00 | 43,7545 |
| N válido (por lista) | 20 | | | |

**Interpretación:** "Se observa que en la tabla 3 obtuvo como nivel de morbilidad un mínimo 7.69% y como máximo 61% de morbilidad y de igual manera se observa que después de la implementación se obtuvo un mínimo de 9.09% y un máximo de 100% de morbilidad".

***Figura 6.***
*Nivel de morbilidad - Antes y después de la implementación*




**Interpretación:** Se observa en la figura 6 que el nivel de morbilidad antes de la implementación fue de 35.13% y después de la implementación fue de 43.75%. Como se puede visualizar, hay un aumento de 8.62% después de la implementación.

**Dimensión 2: Tiempo promedio de diagnostico**

**Análisis Descriptivo**

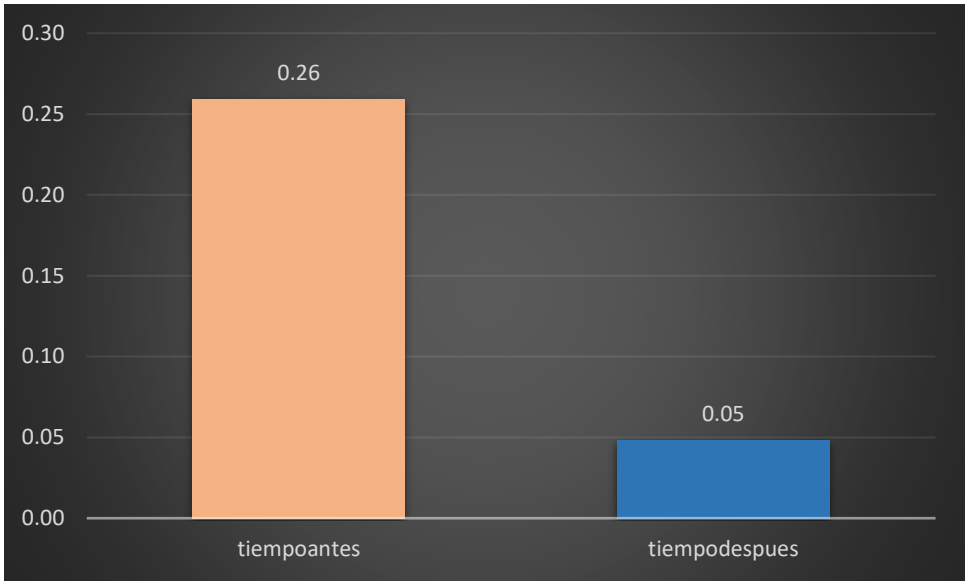
**Tabla 4. Medidas descriptivas de la dimensión 2**

**Estadísticos descriptivos**

| | N | Mínimo | Máximo | Media |
|---|---|---|---|---|
| tiempoantes | 20 | ,12 | ,42 | ,2595 |
| tiempodespues | 20 | ,00 | ,08 | ,0485 |
| N válido (por lista) | 20 | | | |

**Interpretación:** Se observa que en la tabla 4 obtuvo como tiempo promedio un mínimo de 00:12 minutos y como máximo de 00:42 minutos, y de igual manera se observa que después de la implementación se obtuvo un mínimo de 0:00 minutos y un máximo de 0:08 minutos.

**Figura 7. Dimensión de tiempo promedio de diagnostico**




**Interpretación:** Se observa en la figura 7 que el tiempo promedio antes de la implementación fue 00:26 minutos y después de la implementación fue de 00:05 minutos. Como se puede visualizar hay una disminución de 00:21 minutos después de la implementación.

## 4.3. Prueba de Hipótesis

### 4.3.1. Hipótesis especificas 1:

$H_0$: La aplicación de minería de datos **no disminuirá** el nivel de morbilidad de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

$$H_0=Nma - Nmd <=0$$

$H_1$: La aplicación de minería de datos **disminuirá** el nivel de morbilidad de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

$$H_1=Nma - Nmd > 0$$

**Tabla 5. Prueba de normalidad de la Dimensión 1**

**Pruebas de normalidad**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| Resultado | ,204 | 20 | ,029 | ,920 | 20 | ,009 |

a. Corrección de significación de Lilliefors

**Interpretación:** Se observa en la tabla 5 que la significancia es de valor 0.009, cuyo valor es menor que 0.05, es por ello que los datos no cumplen o siguen una distribución normal, por eso se optó utilizar una prueba no paramétrica, la cual fue Wilcoxon por ser una población menor a 30.

**Tabla 6.**

**Prueba de rangos con signo de Wilcoxon de la Dimensión 1**

**Estadísticos de prueba[a]**

| | Despues - Antes |
|---|---|
| Z | -3,096[b] |
| Sig. asintótica(bilateral) | ,002 |

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos negativos.

**Tabla 7. Prueba Z de la Dimensión 1**

**Rangos**

| | | N | Rango promedio | Suma de rangos |
|---|---|---|---|---|
| Despues - Antes | Rangos negativos | 2[a] | 11,25 | 22,50 |
| | Rangos positivos | 18[b] | 10,42 | 187,50 |
| | Empates | 0[c] | | |
| | Total | 20 | | |

a. Despues < Antes

b. Despues > Antes

c. Despues = Antes

**Interpretación:** Se acepta la hipótesis alterna con un 95% de confianza, donde la aplicación de minería de datos disminuirá el nivel de morbilidad de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023, dado que Z = -3.096 así como p (Sig.) es mayor que 0.05(0.002 < 0.05) y se rechaza la hipótesis nula.

### 4.3.2. Hipótesis especificas 2:

$H_0$: La aplicación de minería de datos **no disminuirá** el tiempo promedio de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

$$H_0 = Tpa - Tpd <= 0$$

$H_1$: La aplicación de minería de datos **disminuirá** el tiempo promedio de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.

$$H_1 = Tpa - Tpd > 0$$

**Tabla 8. Prueba de normalidad de la Dimensión 2**

**Pruebas de normalidad**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Estadístico | gl | Sig. | Estadístico | gl | Sig. |
| Resultado2 | ,164 | 20 | ,165 | ,928 | 20 | ,002 |

a. Corrección de significación de Lilliefors

**Interpretación:** Se observa en la tabla 8 que la significancia es de valor 0.002, cuyo valor es menor que 0.05, es por ello que los datos no cumplen o siguen una distribución normal, por eso se optó utilizar una prueba no paramétrica, la cual fue Wilcoxon por ser una población menor a 30.

**Tabla 9. Prueba de rangos con signo de Wilcoxon de la Dimensión 2**

**Rangos**

| | | N | Rango promedio | Suma de rangos |
|---|---|---|---|---|
| tiempodespues - tiempoantes | Rangos negativos | 20[a] | 10,50 | 210,00 |
| | Rangos positivos | 0[b] | ,00 | ,00 |
| | Empates | 0[c] | | |
| | Total | 20 | | |

a. tiempodespues < tiempoantes

b. tiempodespues > tiempoantes

c. tiempodespues = tiempoantes

**Tabla 10. Prueba de normalidad de la Dimensión 2**

**Estadísticos de prueba[a]**

| | tiempodespues - tiempoantes |
|---|---|
| Z | -3,927[b] |
| Sig. asintótica(bilateral) | ,000 |

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos positivos.

**Interpretación:** Se acepta la hipótesis alterna con un 95% de confianza, la aplicación de minería de datos disminuirá el tiempo promedio de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023, dado que Z = -3.927 así como p (Sig.) es menor que 0.05 (0.000 < 0.05) y se rechaza la hipótesis nula.

## 4.4. Discusión de resultados

Ante el resultado obtenido por mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, al implementar una aplicación de minera de datos donde, una de sus efectividades fue que se pudo disminuir el tiempo promedio necesario para reconocer la existencia de enfermedades respiratorias.

Con el primer indicador del nivel de morbilidad, se contaba con un 35.13% en el nivel de morbilidad antes de la implementación siendo este un nivel negativo, posterior a ello con la aplicación ya implementada el nivel aumento negativamente a un 43.75%, estos datos son comparados a base del estudio de Cordova et al. (2020), en el cual determina un nivel de 30.47% según los casos reportados en niños del centro materno infantil, declarando un índice altamente negativo.

Con el indicador tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias, previamente se contaba con 00:25

minutos siendo este un tiempo alto y después con 00:05 minutos. Se logró tener una reducción de 00:21 minutos, estos datos son respaldados por el estudio de Marimón y Navarro (2017), en el cual determina un tiempo de diagnóstico entre 15 a 30 min y cómo influye directamente en la detección de enfermedades a base de los antecedentes del paciente, declarándolo como tiempo deficiente.

## CONCLUSIONES

- Se concluyó que se mejoraría la identificación de enfermedades respiratorias implementando una aplicación de minería de datos en el Hospital Daniel Alcides Carrión Pasco, 2023. Al disminuir el tiempo promedio.
- Aseguró que la frecuencia no bajó significativamente por debajo del nivel de enfermedad, Antes de implementar la aplicación de minería de datos obtuvimos un resultado de 35,13% (positivo) y después de implementar la aplicación de minería de datos obtuvimos un resultado de 43,75% (negativo). Aumento del 8,62%. La prueba estadística de Wilcoxon muestra que la puntuación Z es -3,096 y siguientes con un 5% de significancia y un 95% de confianza. 0.002.
- Determinar la reducción del tiempo promedio para detectar la presencia de complejos de enfermedades respiratorias. Los resultados se obtienen 00:26 minutos antes y 00:05 minutos después de implementar la aplicación de minería de datos, lo que resulta en un ahorro de 00:21 minutos. El nivel de significancia al nivel de confianza del 5% y al nivel del 95% utilizando el estadístico de prueba de Wilcoxon y la puntuación Z es -3,927 con signo 0,000.

## RECOMENDACIONES

- El personal está en mejores condiciones de monitorear a los pacientes para tomar medidas preventivas que eviten la recurrencia de la enfermedad y brindar recomendaciones oportunas para futuros exámenes a intervalos regulares, lo cual está planificado.
- Recomendar al personal administrativo y técnico de la oficina a organizar los datos de los pacientes de manera centralizada, especialmente para los registros médicos de más de un año, para reducir el tiempo dedicado a recuperar archivos (registros médicos) de diversas fuentes.

## REFERENCIAS BIBLIOGRÁFICAS


Alva Mariños, R. S., & Cruz Isla, L. F. (2021). Aplicación de minería de datos para mejorar el diagnóstico de un grupo de enfermedades respiratorias en un hospital de Trujillo [UNIVERSIDAD CÉSAR VALLEJO]. In *UNIVERSIDAD CÉSAR VALLEJO*. http://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/47102/Gutierrez_RS-SD.pdf?sequence=1&isAllowed=y

Bautista, L. (2010). *Uso de minería de datos en la detección temprana y prevención de complicaciones de enfermedades en el sistema de salud Colombiano* (Vol. 2, Issue 5). UNIVERSIDAD DE LOS ANDES BOGOTÁ D.C.

Ccopa Mamani, M., & Chavez Viza, S. V. (2015). Modelo Predictivo Basado en Minería de Datos Para la Mejora en la Toma de decisiones del departamento de Cirugía del Hospital Regional Manuel Núñez Butrón [UNIVERSIDAD NACIONAL DEL ALTIPLANO]. In *Universidad Nacional del Altiplano*. http://repositorio.unap.edu.pe/handle/UNAP/1911

Curi, H. (2020). Reconocimiento de patrones en enfermedades respiratorias mediante minería de datos para mejorar el diagnostico en pacientes del Hospital Daniel Alcides Carrión – Pasco [UNIVERSIDAD DANIEL ALCIDES CARRIÓN]. In *Interciencia* (Vol. 1, Issue 1). http://repositorio.usanpedro.edu.pe/bitstream/handle/USANPEDRO/6050/Tesis_57389.pdf?sequence=1&isAllowed=y%0Ahttp://cybertesis.unmsm.edu.pe/handle/cybertesis/10302%0Ahttp://repositorio.undac.edu.pe/bitstream/undac/414/1/T026_70261078_T.pdf

Hernández Sampieri, R., Férnandez Collado, C., & Baptista Lucio, P. (2018). *Metodología de la Investigación* (S. A. D. C. . McGRAW-HILL / INTERAMERICANA EDITORES (ed.); Sexta).

Hernández Sampieri, R., & Mendoza Torres, C. (2018). *Metodología de la Investigación*.

Jara Tucto, A. (2023). *Identificación automática de neumonía mediante el procesamiento digital del sonido*. UNIVERSIDAD SEÑOR DE SIPÁN.

Jimenéz, M. (2019). *Aplicación de analítica de datos para predicción de infección respiratoria aguda en Colombia*. Universidad de los Andes.

Ortega Rojas, Y. (2021). *"USO DE LAS HERRAMIENTAS TECNOLÓGICAS – TIC Y SU INFLUENCIA EN EL RENDIMIENTO ACADÉMICO EN LOS ESTUDIANTES DE LA ESCUELA PROFESIONAL DE ENFERMERÍA UNAC-2020."* UNIVERSIDAD NACIONAL DEL CALLAO.

Oviedo, E., Oviedo, A., & Vélez, G. (2023). Minería de Datos: Aportes y tendencias en el Servicio de Salud de Ciudades Inteligentes. *Revista Politécnica*, *52*(1), 111–120. https://doi.org/10.33333/rp.vol52n1

Rojas, E., & Sebastián, J. (2017). Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia [UNIVERSIDAD CATÓLICA DE COLOMBIA]. In *Universitas Nusantara PGRI Kediri* (Vol. 01). http://www.albayan.ae

**ANEXOS**

**Instrumentos de Recolección de datos**

| Ficha de Registro | | | |
|---|---|---|---|
| **Investigadores** | | **Tipo de prueba** | |
| **Empresa investigada** | Hospital Daniel Alcides Carrión Pasco | | |
| **Motivo de investigación** | Nivel de Morbilidad | | |
| **Fecha de inicio** | | **Fecha de termino** | |
| **Objetivo** | **Indicador** | **Medida** | **Formula** |
| Disminuir el nivel de morbilidad | Nivel de morbilidad | Porcentaje | $NM = (TC\ x\ 100)\ /\ TCE$ |

| N° | Paciente | Procedencia del paciente | RAM (Reacciones alérgicas al medicamento) | Fecha de nacimiento | Fecha de diagnóstico | Enfermedad diagnosticada |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | | | | | | |
| 17 | | | | | | |
| 18 | | | | | | |
| 19 | | | | | | |

| Ficha de Registro | | | |
|---|---|---|---|
| **Investigadores** | | **Tipo de prueba** | |
| **Empresa investigada** | Hospital Daniel Alcides Carrión Pasco | | |
| **Motivo de investigación** | Tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias | | |
| **Fecha de inicio** | | **Fecha de termino** | |
| **Objetivo** | **Indicador** | **Medida** | **Formula** |
| Disminuir el tiempo promedio | Tiempo promedio | Horas | $TPIEGER = \frac{\sum_{i=1}^{n}(TIEER)i}{n}$ |

| N° | Paciente | Procedencia del paciente | RAM (Reacciones alérgicas al medicamento) | Fecha y Hora Ingreso (dd/mm/yyyy - hh:mm) | Fecha - Hora Salida (dd/mm/yyyy - hh:mm) | Fecha y Hora de diagnóstico (dd/mm/yyyy - hh:mm) | Enfermedad diagnosticada | Numero de análisis adicionales | Días de tratamiento farmacológico |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | | | | | | | | | |
| 5 | | | | | | | | | |
| 6 | | | | | | | | | |
| 7 | | | | | | | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | |
| 17 | | | | | | | | | |
| 18 | | | | | | | | | |
| 19 | | | | | | | | | |

Desarrollo de la Metodología

**Fase 1: Comprensión del negocio**

**Determinar los objetivos del negocio**

- Recopilar información de la empresa

En el Hospital Daniel Alcides Carrión Pasco se detectó un incremento de casos, por lo que, en la mayoría de estos fueron enfermedades respiratorias, los cuales afectaron a los adultos, viéndose reflejado la poca eficacia en el proceso de diagnóstico médico y tratamientos en el departamento de Neumología de dicho conjunto de enfermedades, ocasionando un aumento en el nivel de morbilidad, tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias y costo promedio de diagnóstico del dicho grupo.

- Definir los objetivos del negocio:

Brindar una atención con calidad y calidez a miles de asegurados.

Atender a pacientes con enfermedades de especialidades complejas.

- Definir los criterios de rendimiento comercial:

La posibilidad de realizar diagnósticos a pacientes con enfermedades respiratorias con un elevado porcentaje de fiabilidad.

**Evaluación de la situación**

- Definir requisitos

Reporte de Historias clinicas de pacientes

Reporte de costos

Reporte de atenciones medicas

- Definir riesgos

Acceso a la base de datos

- Definir contingencias

Solicitar reportes de historias clínicas

**Determinar los objetivos de minería de datos**

- Definir objetivos de mineria de datos

✓ Nivel de morbilidad

✓ Disminuir el tiempo promedio para identificar la existencia de un grupo de enfermedades respiratorias.

✓ Disminuir el costo promedio de diagnóstico de un grupo de enfermedades respiratorias.

- Definir criterios de rendimiento de minería de datos

Se establece como criterio la posibilidad de realizar predicciones sobre la enfermedad respiratoria que sufre un paciente con un elevado porcentaje de fiabilidad.

**Realizar el plan del proyecto**

**Fase 2: Comprensión de los datos**

**Recolectar los datos iniciales**

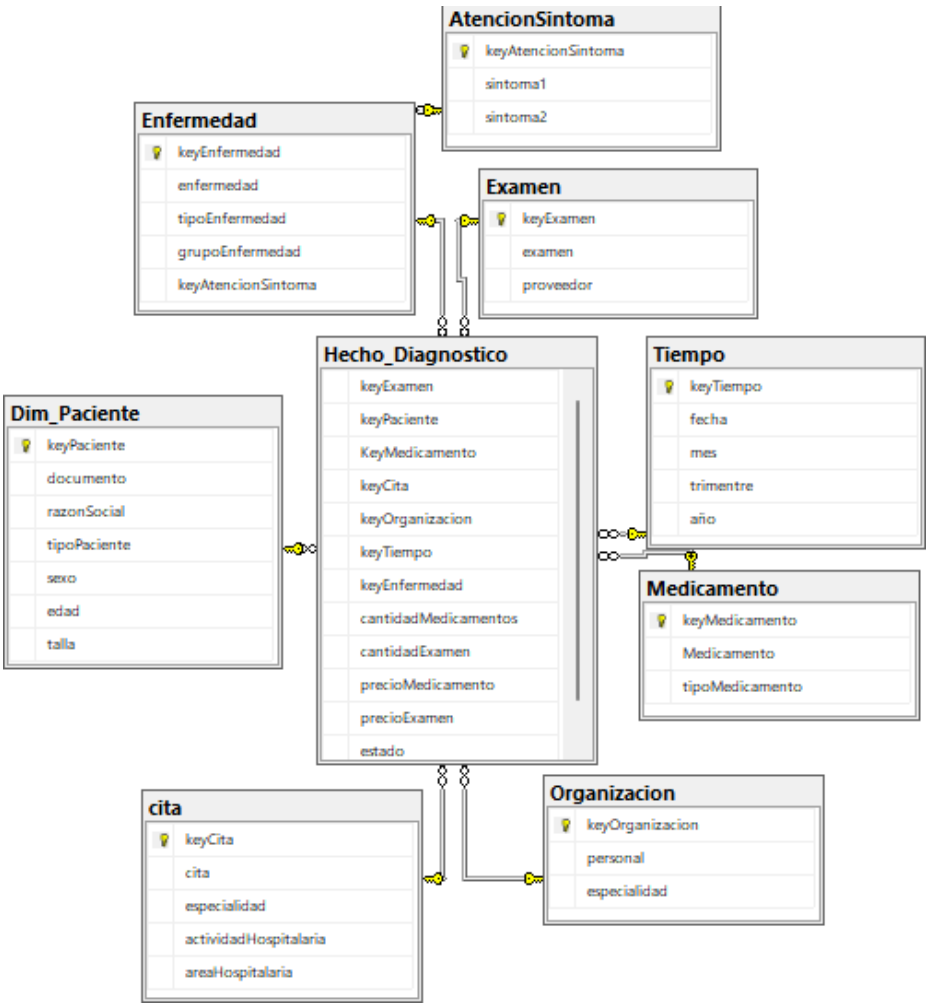
Los datos empleados en este proyecto de investigación son de historias medicas de pacientes del Hospital Daniel Alcides Carrión Pasco que incluyen información personal sobre ellos como: nombres, apellidos, DNI, fecha de nacimiento, entre otros.

Listado de los datos adquiridos:

- Paciente
- Enfermedad
- Cita
- Examen
- Medicamento
- Personal

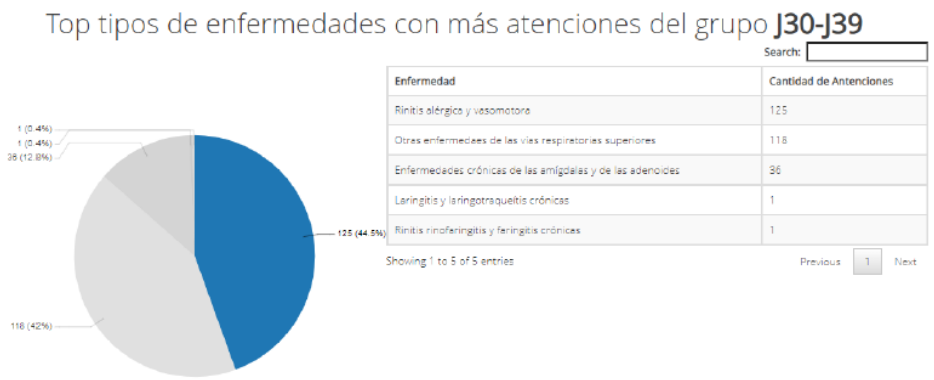
## Descripción de los datos

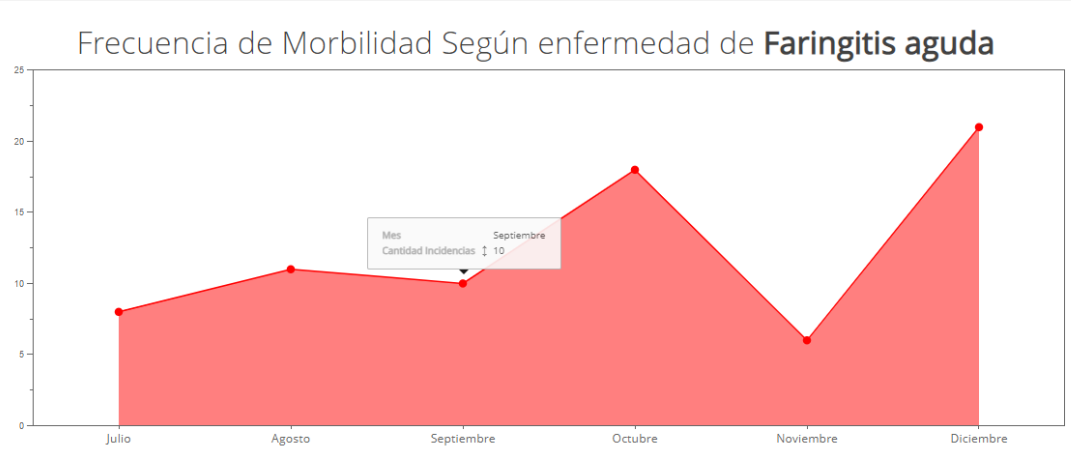
Los datos se encuentran almacenados en un almacén de datos con esquema dimensional en estrella. Podemos observar el esquema relaciona de la base de datos, la cual fue generada por la herramienta SQL.




## Exploración de los datos

Una vez que se han descrito los datos, se procede a explorarlos, esto implica aplicar pruebas estadísticas básicas que revelarán propiedades de los datos, y crear tablas de frecuencia y gráficos de distribución de los datos. Este informe sirve principalmente para determinar la consistencia y completitud de los datos.




**Verificar la calidad de los datos**

Después de hacer la exploración inicial de los datos se puede afirmar que estos son completos. Los datos cubren los casos requeridos para la obtención de los resultados necesarios para poder cumplir los objetivos del proyecto. Los datos no contienen errores, estos son datos propios de las actas médicas. Tampoco se encuentran valores fuera de rango, dado que los datos son controlados por medio de la integración, por lo que no hay riesgo de ruido en el proceso de la minería de datos.

**Fase 3: Preparación de los datos**

**Seleccionar los datos**

En términos de registros, se van a utilizar todos los registros dentro de cada tabla que compone la base de datos, puesto que al ser ésta una base de datos específicamente creada para este proyecto. Sin embargo, hay campos dentro de

estos registros que no son necesarios para nuestros objetivos de minería de datos, por lo que se puede prescindir de algunos de ellos.

Tablas a usar

- AtencionSintoma
- Enfermedad
- Hecho_Diagnostico

**Limpiar los datos**

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir los objetivos, además, estos datos son obtenidos por atenciones médicas para el caso que se presenta, son datos limpios y por lo tanto no hay necesidad de hacer una limpieza más profunda sobre ellos.

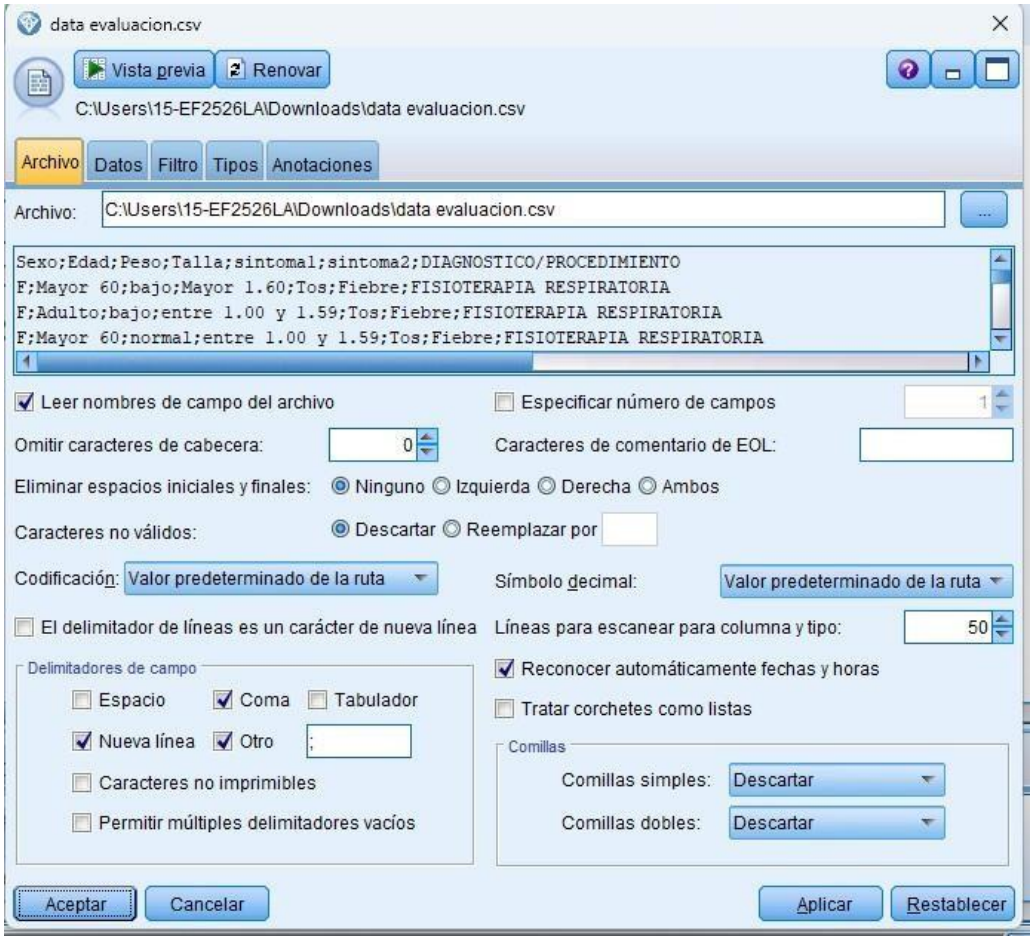
Tampoco tenemos campos en los que falten valores, más allá de los valores nulos que aparecen cuando la información que se quiere representar no existe, y por lo tanto no se consideran como datos faltantes, por lo que no es necesario realizar ningún tipo de estimación de valores faltantes. Estos valores nulos se tratarán a la hora de hacer la minería de datos simplemente ignorándolos porque no aportan ninguna información adicional al estudio.

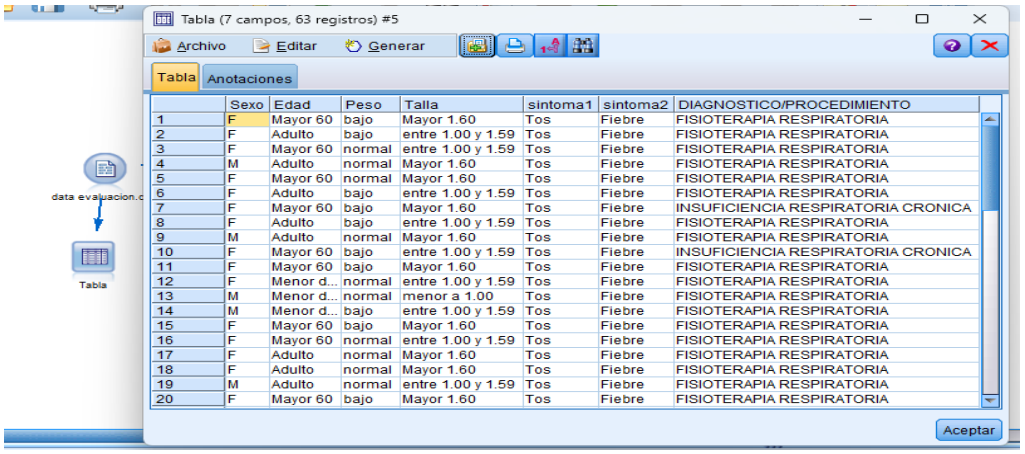
Posteriormente, se da importancia a los datos que conforman el modelo a construir. Para esto, se usa la tabla de hechos que contiene la información original de las variables que se predecirán y que serán utilizadas para generar el modelo predictivo.

| Sexo | Edad | Peso | Talla | sintoma1 | sintoma2 | DIAGNOSTICO/PROCEDIMIENTO |
|---|---|---|---|---|---|---|
| F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Adulto | bajo | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | normal | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| M | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Adulto | bajo | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | INSUFICIENCIA RESPIRATORIA CRONICA |
| F | Adulto | bajo | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| M | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | bajo | entre 1.00 y 1.5 | Tos | Fiebre | INSUFICIENCIA RESPIRATORIA CRONICA |
| F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Menor de 18 | normal | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| M | Menor de 18 | normal | menor a 1.00 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| M | Menor de 18 | bajo | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Mayor 60 | normal | entre 1.00 y 1.5 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| F | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |

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

A continuación, se procede a trabajar en el IBM SPSS Modeler versión 18.0, importando la tabla de hechos ya creada y optimizada, llamada dataevaluacion.csv, como se muestra en el siguiente paso:




| | Sexo | Edad | Peso | Talla | sintoma1 | sintoma2 | DIAGNOSTICO/PROCEDIMIENTO |
|---|---|---|---|---|---|---|---|
| 1 | F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 2 | F | Adulto | bajo | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 3 | F | Mayor 60 | normal | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 4 | M | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 5 | F | Mayor 60 | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 6 | F | Adulto | bajo | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 7 | F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | INSUFICIENCIA RESPIRATORIA CRONICA |
| 8 | F | Adulto | bajo | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 9 | M | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 10 | F | Mayor 60 | bajo | entre 1.00 y 1.59 | Tos | Fiebre | INSUFICIENCIA RESPIRATORIA CRONICA |
| 11 | F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 12 | F | Menor d... | normal | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 13 | M | Menor d... | normal | menor a 1.00 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 14 | M | Menor d... | bajo | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 15 | F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 16 | F | Mayor 60 | normal | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 17 | F | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 18 | F | Adulto | normal | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 19 | M | Adulto | normal | entre 1.00 y 1.59 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |
| 20 | F | Mayor 60 | bajo | Mayor 1.60 | Tos | Fiebre | FISIOTERAPIA RESPIRATORIA |

Importación de la tabla de hechos a IBM SPSS Modeler.

Una vez importada la data, se procede a seleccionar los elementos o variables predictoras, que serán catalogadas como entradas, así como la variable resultado o salida.


Tipos de variables declaradas

**Fase 4: Modelado**

**Escoger la Técnica de Modelado**

Después de deliberar, se decidió utilizar la técnica de modelado basada en clasificación, y se eligieron tres algoritmos para realizar pruebas.

- El algoritmo de Árbol de Clasificación y Regresión (C&R): Construye un árbol de decisión que puede predecir o clasificar observaciones futuras. Este método utiliza la partición iterativa de los datos de entrenamiento para minimizar la impureza en cada paso, buscando alcanzar nodos "puros" donde el 100% de los casos pertenezcan a una categoría específica del objetivo. Los atributos de entrada y el objetivo pueden ser tanto numéricos como categóricos. Cada división en el árbol es binaria, generando dos subgrupos.

- El algoritmo C5.0: también construye un árbol de decisión o un conjunto de reglas, seleccionando el atributo que proporciona la máxima ganancia de información en cada nivel. Este método requiere que el atributo objetivo sea categórico y permite múltiples divisiones, generando más de dos subgrupos.
- El algoritmo CHAID: construye árboles de decisión utilizando estadísticas de chi-cuadrado para identificar las divisiones óptimas. A diferencia de C&R y C5.0, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones pueden generar más de dos ramas. Al igual que los otros algoritmos, puede manejar tanto atributos numéricos como categóricos. CHAID exhaustivo es una variante que examina todas las posibles divisiones con mayor precisión, aunque requiere más tiempo de cálculo.
- Construir el modelo: Usando los tres algoritmos de árboles de decisión mencionados, se desarrollaron modelos predictivos en la investigación. Estos modelos se entrenaron utilizando los 63 registros de datos, lo que resultó en los siguientes niveles predictivos de diagnóstico de neumonía basados en las variables descriptivas:
- Algoritmo árbol de clasificación y regresión (C&R)

| | | |
|---|---|---|
| **Correctos** | 54 | 85,71% |
| **Erróneos** | 9 | 14,29% |
| **Total** | 63 | |

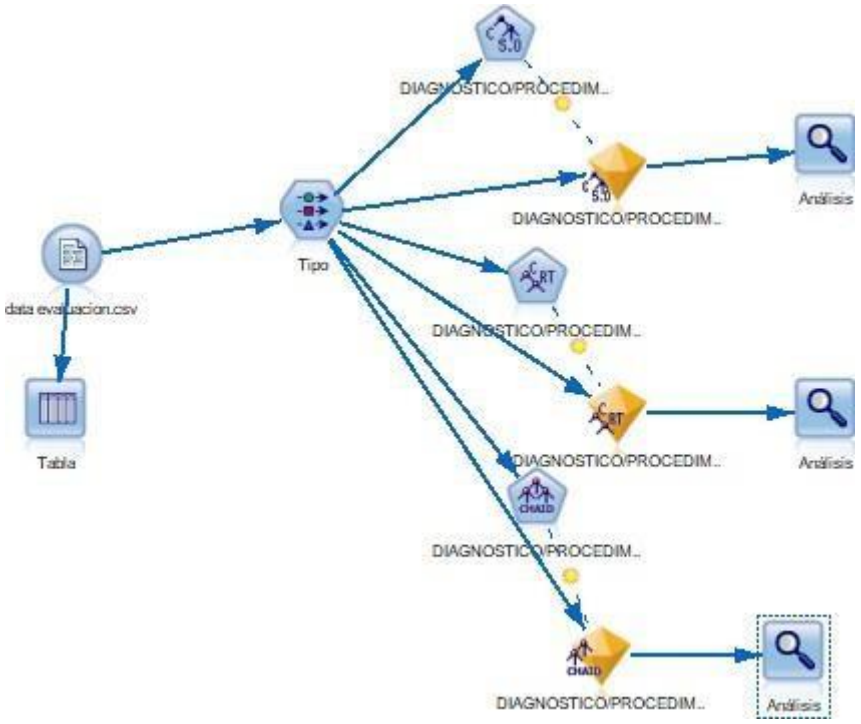
Algoritmo Nodo C5.0

| | | |
|---|---|---|
| **Correctos** | 51 | 80,95% |
| **Erróneos** | 12 | 19,05% |
| **Total** | 63 | |

Algoritmo nodo CHAID

| | | |
|---|---|---|
| **Correctos** | 52 | 82,54% |
| **Erróneos** | 11 | 17,46% |
| **Total** | 63 | |

Se observa que el algoritmo C&R resulta ser el más eficiente durante el entrenamiento, logrando un éxito predictivo del 85.71% en todos los casos de diagnóstico de enfermedades respiratorias. Por lo tanto, se considera el más adecuado para esta investigación.




Aplicación de algoritmos de árboles de decisión

**Fase 5: Evaluación**

En esta fase de la metodología se intentan evaluar los modelos generados, los resultados obtenidos de la aplicación de minería de datos en los patrones de enfermedades respiratorias son las siguientes:

A continuación, se presenta el árbol de decisión, que muestra las variables predictoras que influyen en el diagnóstico de enfermedades respiratorias.

Modelo predictivo algoritmo árbol de decisión

Según el árbol de decisión obtenido:

- La primera división se basa en la variable "Edad", donde se divide en dos grupos: "Adulto" y "Mayor de 60 / Menor de 18". Si la edad es "Adulto", el diagnóstico predicho es "fisioterapia respiratoria".
- Si la edad es "Mayor de 60 / Menor de 18", el árbol se bifurca aún más dependiendo del valor de la variable "Peso". Si el peso es "bajo", el siguiente criterio de división es el "Sexo". Si el sexo es "F", el diagnóstico predicho es "fisioterapia respiratoria".
- Si el sexo es "M" y el peso es "bajo", el árbol se ramifica nuevamente según el valor de la variable "Edad". Si la edad es "Mayor de 60", el diagnóstico predicho es "tbc respiratoria no especificada / tbc pulmonar sin baciloscopia".
- Si el peso es "normal", el diagnóstico predicho es "fisioterapia respiratoria".

**Fase 6: Implantación.**

Esta es la última fase de la metodología CRISP-DM y el objetivo de la misma es el de explicar al cliente como poner en funcionamiento el proyecto que se ha construido en las fases anteriores, así como exponer los resultados obtenidos al cliente de forma que lo pueda entender fácilmente. Otro objetivo de esta fase es el de crear una estrategia para el mantenimiento del proyecto y producir un informe en el que se incluyan posibles mejoras para el futuro y un listado de las dificultades encontradas a la hora de realizarlo.

**Planear la Monitorización y Mantenimiento**

La supervisión y mantenimiento de la implementación del presente proyecto es una fase importante del mismo debido a que los datos que se procesan con mucha frecuencia pueden ser modificados por el personal médico. Los datos pueden ser modificados por diferentes motivos como haber realizado una codificación incorrecta, haber asignado un diagnostico incorrecto al paciente, etc. El volumen de estos datos en movimiento es grande motivo por el cual la extracción de las muestras debe ser

realizada cuidadosamente y realizando siempre backups de los datos explotados en cada proceso. La minería de datos debería ser realizada en periodos mensual, sin embargo, esta medida podría variar en cualquier momento en función de la necesidad que esté vigente en cada momento.

Matriz de Consistencia

**Tema:** “Aplicación de minería de datos para mejorar la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023”

| PROBLEMA GENERAL | OBJETIVO GENERAL | HIPÓTESIS GENERAL | VARIABLE INDEPENDIENTE | DIMENSIÓN | DISEÑO | POBLACIÓN Y MUESTRA |
|---|---|---|---|---|---|---|
| ¿Se podrá aplicar la minería de datos para la mejora en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023? | Aplicar la minería de datos para la mejora en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023. | La aplicación de minería de datos mejora la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023. | Aplicación de minería de datos. | - Análisis | **Diseño:**<br><br>Pre - Experimental<br><br>**Tipo de Investigación**<br><br>Aplicada | **POBLACIÓN**<br><br>La investigación abordada presenta a 56 pacientes<br><br>**MUESTRA**<br><br>La muestra es 20 pacientes mediante la técnica llamada “muestreo por conveniencia” dentro del muestreo no probabilístico. |

| PROBLEMA ESPECÍFICO | OBJETIVO ESPECÍFICO | HIPÓTESIS ESPECIFICA | VARIABLE DEPENDIENTE | DIMENSIÓN | MÉTODO DE INVESTIGACIÓN | TÉCNICAS - INSTRUMENTOS |
|---|---|---|---|---|---|---|
| ¿Se podrá disminuir el nivel de morbilidad aplicando la minería de datos en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023? | Disminuir el nivel de morbilidad aplicando la minería de datos en la identificación de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.<br><br>Disminuir el tiempo promedio aplicando la | La aplicación de minería de datos disminuirá el nivel de morbilidad de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023.<br><br>La aplicación de minería de datos disminuirá el | Identificación de enfermedades respiratorias. | - Nivel de morbilidad<br>- Tiempo promedio de diagnostico | **Método**<br><br>Analítica.<br><br>**Enfoque**<br><br>Cuantitativo | **Técnicas:**<br><br>- Ficha de registro. |

| | | |
|---|---|---|
| ¿Se podrá disminuir el tiempo promedio aplicando la minería de datos para identificar la existencia de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023? | minería de datos para identificar la existencia de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023. | tiempo promedio de enfermedades respiratorias en el Hospital Daniel Alcides Carrión Pasco, 2023. |