ANÁLISIS DE INFORMACIÓN DE UNA FUENTE INTERNACIONAL SOBRE EL COVID-19 PARA ANTICIPAR PROBLEMAS POTENCIALES EN COLOMBIA

Cristian Leandro Botina Caipe, ✉ cristian_botina@hotmail.com
Oscar Eduardo Daza Florez, ✉ oscardf98@gmail.com

Monografía elaborada para optar por el título de Especialista en Gestión de Bases de Datos

Asesor: Ing. Beatriz E. Marin
Coasesores: Nilson Mossos y Juan Carlos Pazmiño

Institución Universitaria Antonio José Camacho
Facultad de Ingenierías
Especialización en Gestión de Bases de Datos
Cali - Colombia
2020

Nota de aceptación:

Aprobado por el Comité de Grado en cumplimiento de los requisitos exigidos por la Institución Universitaria Antonio José Camacho para optar al título de Especialista en Gestión de Bases de Datos

Jurado

Jurado

Santiago de Cali, 17 de junio del 2020

Dedicatoria

Dedicamos este proyecto de grado a Dios, a nuestras familias, quienes fueron el motor moral incondicional para que nuestros esfuerzos dieran frutos.

A nuestros compañeros, amigos del ayer y del hoy, que sin esperar nada a cambio compartieron sus conocimientos en el recorrido de estos años de carrera.

Gracias a todos.

Agradecimientos

En primera instancia agradecer a Dios por regalarnos el Don de la vida

Agradecimiento especial a nuestras familias (padres, hermanos, esposa, hijos) por su apoyo incondicional a lo largo del camino de nuestras carreras.

Reconocimiento a cada docente que aportó un granito de arena con su experiencia y conocimientos; y que resultado de ello se ha llegado hasta ésta instancia.

Tabla de contenido

Comentado [BEMO1]: Todas las subsecciones deben ir numeradas

Lista de tablas

Lista de figuras

Resumen

En este documento se presenta el desarrollo de una solución a nivel de base de datos en MaríaDB con los estándares de seguridad necesarios que sirva de herramienta para recopilar información relacionada al COVID 19 en Colombia. Dicha solución estará soportada en la investigación y análisis de fuentes de información que están a la vanguardia de la problemática sanitaria que afecta actualmente a todo el mundo. Por consiguiente, se realizarán prácticas de minería de datos a fuentes de datos de otros países para tener un objeto de estudio y poder generar conclusiones sobre patrones en el comportamiento y/o evolución del virus en los pacientes.

Abstract

This document presents the development of a database-level solution in MaríaDB with the necessary security standards that serves as a tool to collect information related to COVID 19 in Colombia. This solution will be supported in the research and analysis of information sources that are at the forefront of the health problems that currently affect everyone. Therefore, data mining practices will be carried out to data sources from other countries to have an object of study and to be able to draw conclusions about patterns in the behavior and / or evolution of the virus in patients.

Comentado [BEMO2]: No se usa

Introducción

En la presente monografía se pretende analizar una fuente de datos externa de la Red Unida para la Exploración e Investigación de Datos COVID expuesta en Kaggle, logrando abstraer la información y el conocimiento necesario para generar una nueva base de datos transaccional que sirva para capturar datos sobre el COVID19 en Colombia.

Por eso se propone como objetivo principal utilizar la recopilación estos datos extranjeros para aprender sobre el comportamiento de la información recolectada y generar un conocimiento que sirva para anticipar problemas potenciales relacionados al virus en nuestro país.

Como solución al problema, en el documento se encuentra el diseño y desarrollo de una base de datos transaccional, junto con la aplicación de seguridad y protección de datos sensibles en la base de datos desarrollada. La aplicación de un proceso de minería de datos que permita hacer una predicción de los datos obtenidos de la fuente externa encontrar interrelación de los datos con el diagnostico positivo o negativo del virus.

1 Planteamiento del problema

Actualmente se encuentran bancos de datos nutridos de información referente al COVID-19 alrededor del mundo(OPS/OMS 2020) con múltiples fuentes expuestas por las comunidades de la salud. Sin embargo, en Colombia (Datos Abiertos del Gobierno-Colombia) no se ha logrado recolectar datos de los pacientes con COVID-19 de la misma manera como se ha hecho en otros países debido a que no se cuenta con una herramienta transaccional soportada bajo una solución de base de datos que permita recopilar esta información y obtener un aprendizaje de los mismos datos de manera adecuada.

Por ello surge la necesidad de construir e implementar una base de datos transaccional bajo estándares de seguridad estipulados por la norma ISO 27000 que permita recopilar información sobre la problemática sanitaria del coronavirus en Colombia. Como estudio complementario se opta por aplicar técnicas de minería de datos a información recopilada en otros países que permita seguir un modelo a implementar en la información que se obtenga a futuro en Colombia.

Es necesario apoyarse en recursos de información como fuentes de datos diversas y oficiales como la OMS y locales que sirvan como estrategia de aprendizaje.

Se pretende poder almacenar y estructurar esta información con el fin de que ésta pueda ser útil para identificar a futuro, modelos o patrones de casos COVID 19 que se serán objeto de estudio en Colombia; por ello se aplicará técnicas de minería de datos a otros conjuntos de datos ya existentes de otros países (Roche Data Science Coalition, 2020) que sirvan de insumo o referencia ante futuros patrones identificados en Colombia.

Comentado [BEMO3]: Si esto es una referencia debe estar bien formulada en norma APA y estar asociado en el capítulo de referencias

2 Objetivos

2.1 Objetivo general

Utilizar la información existente en diversos medios internacionales sobre el COVID-19, para aprender sobre el comportamiento de estos datos recolectados en otros países logrando un aprendizaje que sirva para anticipar problemas potenciales en Colombia.

2.2 Objetivos específicos

- Desarrollar una base de datos que permita guardar la información de estos datos en Colombia
- Implementar el uso de políticas que permitan la seguridad y confiabilidad de la información almacenada y procesada.
- Aplicar una técnica de minería de datos a una fuente de datos externas que permita hallar una tendencia o patrones de los datos recolectados
- Presentar la información obtenida por medio de una herramienta de consulta y análisis.

3 Marco teórico

El Desarrollo del presente trabajo se orientó a la problemática sanitaria que afecta al mundo entero, para lo cual es necesario aclarar en primera instancia, que el término Coronavirus no es equivalente a Covid 19 (Brian y Baric 2005).

3.1 Coronavirus y Covid-19

Los coronavirus son una extensa familia de virus que pueden causar enfermedades tanto en animales como en humanos. En los humanos, se sabe que varios coronavirus causan infecciones respiratorias que pueden ir desde el resfriado común hasta enfermedades más graves como el síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SRAS). El coronavirus que se ha descubierto más recientemente causa la enfermedad por coronavirus COVID-19 (OMS s. f.)

La COVID-19 es la enfermedad infecciosa causada por el coronavirus que se ha descubierto más recientemente. Tanto este nuevo virus como la enfermedad que provoca eran desconocidos antes de que estallara el brote en Wuhan (China) en diciembre de 2019. Actualmente la COVID-19 es una pandemia que afecta a muchos países de todo el mundo(Marra et al. 2003).

3.2 Síntomas de la COVID -19

Los síntomas más habituales de la COVID-19 son la fiebre, la tos seca y el cansancio. Otros síntomas menos frecuentes que afectan a algunos pacientes son los dolores y molestias, la congestión nasal, el dolor de cabeza, la conjuntivitis, el dolor de garganta, la diarrea, la pérdida del gusto o el olfato y las erupciones cutáneas o cambios de color en los dedos de las manos o los pies. Estos síntomas suelen ser leves y comienzan gradualmente. Algunas de las personas infectadas solo presentan síntomas leves(Novel, 2020).

La mayoría de las personas (alrededor del 80%) se recuperan de la enfermedad sin necesidad de tratamiento hospitalario. Alrededor de 1 de cada 5 personas que contraen la COVID-19 acaba presentando un cuadro grave y experimenta dificultades para respirar. Las personas mayores y las que padecen afecciones médicas previas como hipertensión arterial, problemas

cardiacos o pulmonares, diabetes o cáncer tienen más probabilidades de presentar cuadros graves. Sin embargo, cualquier persona puede contraer la COVID-19 y caer gravemente enferma. Las personas de cualquier edad que tengan fiebre o tos y además respiren con dificultad, sientan dolor u opresión en el pecho o tengan dificultades para hablar o moverse deben solicitar atención médica inmediatamente.

3.3 Herramientas de Trabajo

3.3.1 Sistema de Base de Datos

Para el diseño del modelo Lógico y Físico de la base de datos se eligió MariaDB y la herramienta , el cual es un Sistema de Gestión de Bases de Datos derivado de MySQL, una de las bases de datos más importantes del mercado (Bartholomew 2012), utilizado para manejar grandes cantidades de información como por ejemplo Facebook, Twitter y Wikipedia.

La simplicidad de la sintaxis en MariaDB, permite crear bases de datos simples o complejos con mucha facilidad; es compatible con múltiples plataformas informáticas y está provista de una infinidad de aplicaciones que permiten acceder rápidamente a las sentencias de la gestión de base de datos.

3.4 Derecho Habeas Data

El Derecho de Habeas Data consiste en Colombia por permitir a los ciudadanos conocer, actualizar y rectificar toda la información que tengan las diferentes entidades y bases de datos del país. surgió como parte fundamental de los artículos 15 y 20 de la Constitución Política. Posteriormente, se creó como derecho autónomo e independiente, que gozan todos los colombianos (Ley 1266 de 2008).

Para los efectos de la presente ley, se entiende por:

a) Titular de la información. Es la persona natural o jurídica a quien se refiere la información que reposa en un banco de datos y sujeto del derecho de hábeas data y demás derechos y garantías a que se refiere la presente ley;

b) Fuente de información. Es la persona, entidad u organización que recibe o conoce datos personales de los titulares de la información, en virtud de una relación comercial o de servicio o de cualquier otra índole y que, en razón de autorización legal o del titular, suministra esos datos a un operador de información, el que a su vez los entregará al usuario final. Si la fuente entrega la información directamente a los usuarios y no, a través de un operador, aquella tendrá la doble condición de fuente y operador y asumirá los deberes y responsabilidades de ambos. La fuente de la información responde por la calidad de los datos suministrados al operador la cual, en cuanto tiene acceso y suministra información personal de terceros, se sujeta al cumplimiento de los deberes y responsabilidades previstas para garantizar la protección de los derechos del titular de los datos.

c) Operador de información. Se denomina operador de información a la persona, entidad u organización que recibe de la fuente datos personales sobre varios titulares de la información, los administra y los pone en conocimiento de los usuarios bajo los parámetros de la presente ley. Por tanto el operador, en cuanto tiene acceso a información personal de terceros, se sujeta al cumplimiento de los deberes y responsabilidades previstos para garantizar la protección de los derechos del titular de los datos. Salvo que el operador sea la misma fuente de la información, este no tiene relación comercial o de servicio con el titular y por ende no es responsable por la calidad de los datos que le sean suministrados por la fuente;

d) Usuario. El usuario es la persona natural o jurídica que, en los términos y circunstancias previstos en la presente ley, puede acceder a información personal de uno o varios titulares de la información suministrada por el operador o por la fuente, o directamente por el titular de la información. El usuario, en cuanto tiene acceso a información personal de terceros, se sujeta al cumplimiento de los deberes y responsabilidades previstos para garantizar la protección de los derechos del titular de los datos. En el caso en que el usuario a su vez entregue la información directamente a un operador, aquella tendrá la doble condición de usuario y fuente, y asumirá los deberes y responsabilidades de ambos;

e) Dato personal. Es cualquier pieza de información vinculada a una o varias personas determinadas o determinables o que puedan asociarse con una persona natural o jurídica. Los datos personales pueden ser públicos, semiprivados o privados;

f) Dato público. Es el dato calificado como tal según los mandatos de la ley o de la Constitución Política y todos aquellos que no sean semiprivados o privados, de conformidad con la presente ley.

g) Dato semiprivado. Es semiprivado el dato que no tiene naturaleza íntima, reservada, ni pública y cuyo conocimiento o divulgación puede interesar no sólo a su titular sino a cierto sector o grupo de personas o a la sociedad en general.

h) Dato privado. Es el dato que por su naturaleza íntima o reservada sólo es relevante para el titular.

3.5 Seguridad de la Información

Además de las mejoras que tiene MariaDB, permite aplicar en la estructura de la base de datos, buenas prácticas de protección de datos, siguiendo como referencia, Políticas de seguridad previamente establecidas de acuerdo la normatividad que invita a seguir la familia de la norma ISO 27000 y sus derivados (Disterer, 2013).

Al aplicar buenas prácticas de seguridad se garantiza la protección, confidencialidad, integridad y disponibilidad de la información sensible frente a la manipulación por parte de usuarios no autorizados mediante el uso de autenticación soportado en roles y privilegios que limiten el acceso a los objetos y las acciones permitidas en la base de datos. Los datos deben ser sometidos a procesos de auditorías para supervisar los diferentes movimientos en la base de datos de tal manera que pueda descubrirse cualquier acción indebida o errónea (López,2018)

Comentado [BEMO4]: Debe ir de acuerdo a la norma Apa

3.6 Aplicación de Minería de Datos

Para llevar a cabo el objetivo de este trabajo, se aplicará minería de datos usando el paquete de software Weka.Weka es un acrónimo de WaikatoEnvironment for Knowledge Analysis, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario

WEKA se distribuye como software de libre distribución desarrollado en Java. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. (Sierra, 2006)

Comentado [BEMO5]: Aquí solo van apellido y año, en el capítulo de referencias va apellido, primera letra del nombre y año. Aplica para que revise en todas las referencias

3.6.1 Árbol de Decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos: Algoritmo J48, Algoritmo C4.5.(Corso, 2020)

Comentado [BEMO6]: Lopez

3.6.2 Algoritmo de Clasificación

Se considera aplicar algoritmo de Clasificación a los datos de estudio disponible con el objetivo de predecir el resultado de examen de covid 19 en pacientes dado un conjunto de datos relacionados.Dentro del mundo de reconocimiento de patrones existen dos grandes grupos de familias que enfocan de una manera diferente el problema de la clasificación. Por un lado está la clasificación no supervisada, trata a la clasificación como el descubrimiento de las clases de un determinado problema. Es decir que contamos con un conjunto de elementos descriptos por un conjunto de características, sin conocer a que clase pertenece cada uno de ellos. En cambio en la clasificación supervisada enfoca el problema de clasificación de otra manera, es decir parte de un conjunto de elementos descripto por un conjunto de características y conocemos la clase al cual pertenece. A este concepto se lo suele denominar conjunto de datos de entrenamiento o conjunto de aprendizaje. Descripción de Matriz de Confusión. VP (Verdaderos positivos): instancias correctamente reconocidas por el sistema. FN (Falsos negativos): instancias que son positivas y que el sistema dice que no lo son. FP (Falsos positivos): instancias que son negativas pero el sistema dice que no lo es. VN (Verdaderos negativos): instancias que son negativas y correctamente reconocidas como tales. Suponiendo que N es el número del conjunto de datos de entrenamiento, N=VP+FN+FP+VN . El número de instancias clasificadas correctamente es la suma de la

diagonal de la matriz y el resto están clasificadas de forma incorrecta. Otra de las maneras de validar más frecuente es basarnos en la tasa de error y la tasa de acierto de un clasificador. Estas tasas se calculan de la siguiente manera: Tasa de error=FP+FN/N Tasa de acierto: VP+VN/N (Corso,2020)

Comentado [BEMO7]: No cumple APA

4 Resultados

4.1 Diccionario de Datos

El análisis de la información sobre el COVID-19, que maneja actualmente países como Brasil y España (Roche Data Science Coalition s. f.), permitió realizar un análisis investigativo para determinar cuál sería el modelo a diseñar que cumpliera con las necesidades que actualmente existen en Colombia derivada de la deficiente recolección de datos apropiada de los casos en todo el país. Para ello se consolidó el siguiente diccionario de datos:

		METADATOS
CAMPO	DESCRIPCIÓN	ENTIDAD
Identificacion del Paciente	Identificacion del Paciente	PACIENTE
Fecha Nacimiento Paciente	Fecha de Nacimiento del Paciente para conocer su edad	PACIENTE
Nombres Paciente	Primer y segundo nombre del paciente	PACIENTE
Apellidos	Primer y segundo Apellido del paciente	PACIENTE
Direccion	Direccion de residencia paciente	PACIENTE
Telefono Fijo	Telefono Fijo del Paciente	PACIENTE
Telefono Celular	Telefono celular del paciente	PACIENTE
Sexo	Sexo del paciente	PACIENTE
Estrato Social	Estrato socioeconomico del paciente	PACIENTE
Correo electronico	Correo electronico del paciente	PACIENTE
Tipo de Identificación	Tipo de Identificación del paciente	TIPO_IDENTIFICACIÓN
Entidad salud	Entidad de salud a la que esta afiliado el paciente y/o Entidad salud que realizó el diagnóstico	ENTIDAD_SALUD
Tipo de Atencion	Descripción del tipo de Atención que esta recibiendo el paciente	TIPO_ATENCION
Estado	Estado actual del paciente	ESTADO
Fecha Estado	Fecha de reporte de estado del paciente	ESTADO_PACIENTE
Fecha inicio sintomas	Fecha en que iniciaron sintomas al paciente	ESTADO_PACIENTE
Municipio	Municipio donde se encuentra reportado el caso	MUNICIPIO
Departamento	Departamento donde se encuentra reportado el caso	DEPARTAMENTO
Pais	Pais donde se encuentra reportado el caso	PAIS
Nombre Diagnóstico	Descripción de Diagnostico	DIAGNOSTICO
Valor Diagnóstico	Valor resultado examen diagnóstico	DIAGNOSTICO_PACIENTE
Fecha diagnostico	Fecha diagnostico del paciente	DIAGNOSTICO_PACIENTE

Figura 1 Diccionario de datos

4.2 Modelo Entidad Relación

A continuación s representa el modelo de entidad relación MER, donde se realiza una abstracción de la información obtenida desde la fuente internacional y sus datos de Kaggle, permitiendo identificar las respectivas entidades que son objetos del mundo real, en este caso Entidad de Salud, Diagnóstico Paciente, Tipo de Atención, Estado del Paciente, Estado, Pacientes, Tipo de Identificación, Países, Departamentos y Municipios. En este modelo se permite interpretar las relaciones, como lo es que los pacientes presentan estados y los estados conllevan a un tipo de atención, esos estados del paciente ocurren en un lugar o un municipio que pertenece a un departamento y a un país. Los pacientes tienen un tipo de identificación necesarios para su identidad, están afiliados a una EPS y se presentan diagnósticos, los cuales son generados en una entidad de salud. En el modelo se pueden identificar los atributos principales, los valores únicos están subrayados y los valores compuestos están en resaltados.

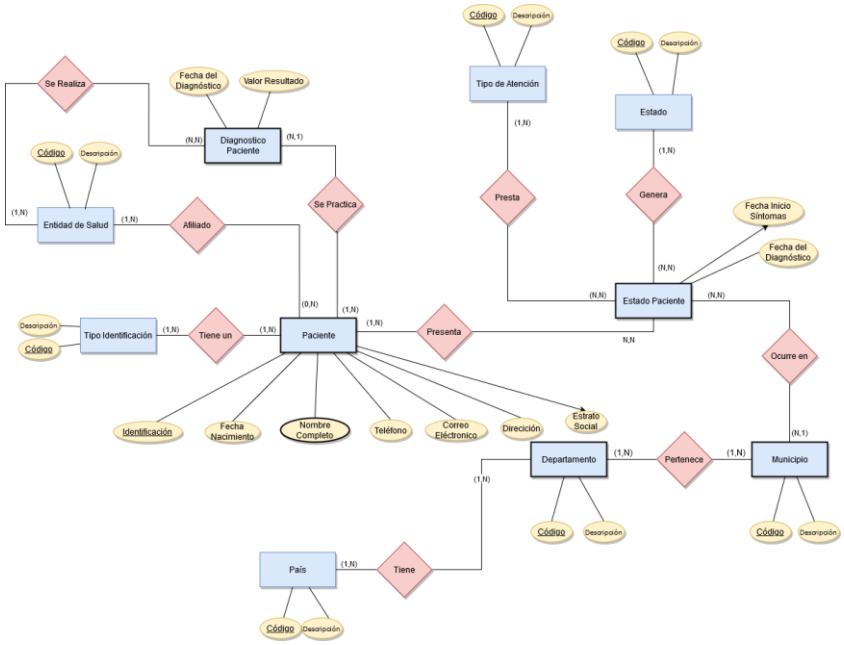

Figura 2 Modelo lógico

4.3 Diseño Físico

A continuación se presenta el diseño obtenido con base en el modelo de entidad relación, lo cual dio como resultado el diseño físico de la base de datos, aquí se pueden visualizar las tablas de la base de datos finales y las relaciones que se reflejan en atributos que se convierten en llaves foráneas permitiendo una integridad referencial.

La herramienta empleada para el modelado fué Mysql Workbench, la cual integra diseño, administración de bases de datos, gestión y mantenimiento para el sistema de bases de datos MariaDB

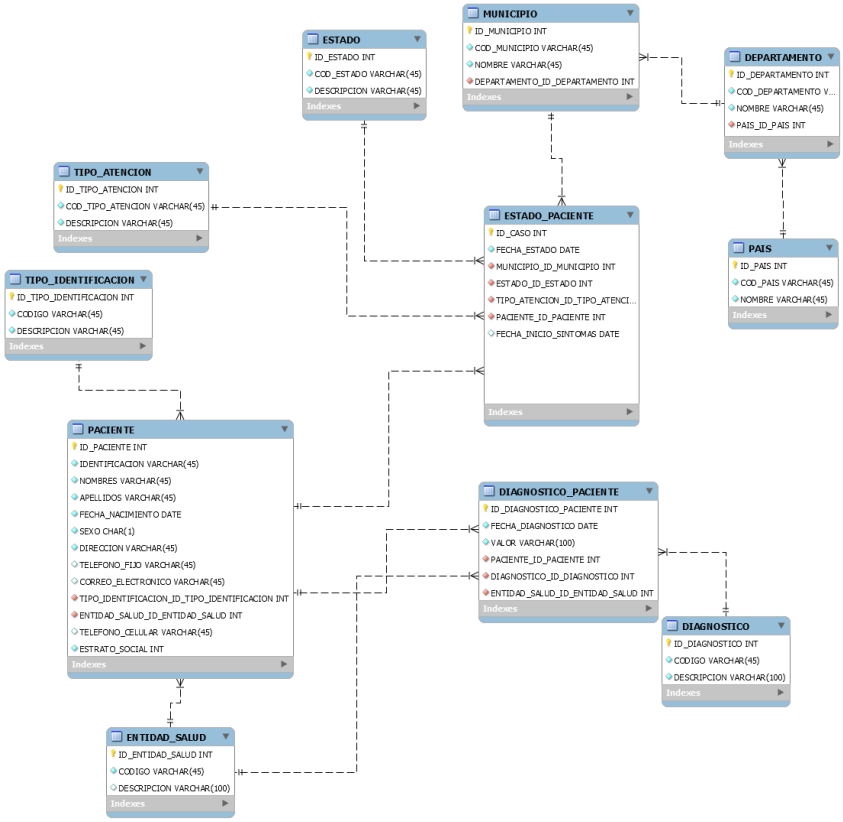

Figura 3MER

4.3 Consultas que satisfacen necesidades del Negocio

El modelo de bases de datos construido satisface las siguientes necesidades:

Comentado [BEMO8]: Organizar bien en cuanto al espacio

- Conocer total de nuevos casos sospechosos de Covid 19 a nivel municipal, departamental y nacional
- Conocer total de nuevos casos positivos de Covid 19 diaria o acumulada a nivel municipal, departamental y nacional
- Conocer total de pacientes fallecidos de manera diaria o acumulada a nivel municipal, departamental y nacional
- Conocer total de pacientes recuperados del virus diaria o acumulada a nivel municipal, departamental y nacional
- Conocer total de pacientes que se encuentran en UCI a nivel municipal, departamental y nacional.
- Conocer total de pacientes cuyo estado de salud es leve a nivel municipal, departamental y nacional.
- Conocer total de pacientes asintomáticos a nivel municipal, departamental y nacional.
- Conocer total de pacientes que se encuentran recibiendo atención en casa a nivel municipal, departamental y nacional.
- Conocer total paciente infectados, recuperados y fallecidos por Estrato Social
- Conocer total pacientes infectados, recuperados y fallecidos por Sexo.
- Conocer total paciente infectados, recuperados y fallecidos por Edad.
- Consultar diagnósticos y/o exámenes aplicados a los pacientes y sus respectivos resultados.

4.4 Seguridad en la Base de Datos

Dando cumplimiento a la Ley de Habeas Data y de acuerdo a lo establecido en la Norma ISO 27001(ISO27000.es s. f.), inicialmente se evaluaron los riesgos que puede presentar la estructura en la base de datos si no se aplica un control que logre reducir o eliminar los problemas de seguridad a los que estaría expuesta la información. Por ello se definieron las siguientes políticas de seguridad:

4.4.1 Organización de seguridad

Definir, coordinar y controlar la gestión necesaria para mitigar los riesgos asociados a la seguridad de la información.

4.4.2 Clasificación y control de activos de información

La información debe estar inventariada y tener identificados los riesgos y exposiciones de seguridad. Los insumos de información deben estar controlados por un responsable que también se encarga de manipularlos de forma correcta.

4.4.3 Tratamiento y gestión del riesgo en seguridad de la información

Analizar los riesgos en seguridad de la información de los pacientes, con base en los objetivos de negocio.

4.4.4 Control de acceso a la información

Implementar las medidas de seguridad aplicables según el caso, con el fin de evitar la adulteración, pérdida, fuga, consulta, uso o acceso no autorizado o fraudulento.

De acuerdo a lo establecido en las políticas de seguridad se desarrollaron las siguientes prácticas sobre la base de datos.

4.4.5 Creación de Usuarios con Roles de TI

USUARIOS Y ROLES DE TI			
Nombre	User	contraseña	Roles
Gilberto Toro Delgado	GilToro	GilToro	DBA
Carlos Hernan Parra Garzon	Carparra	Carparra	MaintenanceAdmin
Michael Yesid Urrea Rojas	MicUrrea	MicUrrea	DBManager
Cristián Leandro Botina Caipe	CriBotina	CriBotina	MonitorAdmin
Oscar Eduardo Daza Florez	OscDaza	OscDaza	ProcessAdmin, BackupAdmin
Cristhian David galvis cardona	CriGalvis	CriGalvis	UserAdmin, DBManager
Eliana Lozano Quenguan	EliQuenguan	EliQuenguan	DBDesigner
Hector Exequiel Rosero Montaño	HecRosero	HecRosero	MaintenanceAdmin

Tabla 1Usuarios y Roles de TI

4.4.6 Creación de usuarios y Roles del sistema

USUARIOS Y ROLES DEL SISTEMA			
Nombre	User	contraseña	Roles
Andrés Gaviria Florez	AGAVIRIAF	AGAVIRIAF	Operador
Alejandra Sánchez Hernández	ASANCHEZH	ASANCHEZH	Operador
Jhon Jairo Florez Gómez	JFLOREZG	JFLOREZ	Usuario
Carlos Andrés Jiménez Castro	CJIMENEZC	CJIMENEZC	Auditor

Tabla 2Usuarios y Roles del Sistema

4.4.7 Encriptado de Información

Con el fin de proteger información sensible asociada al paciente, se construyeron triggers que se disparen al momento de realizar el pre-insert o pre-update de la tabla: **paciente,** con el propósito para encriptar los valores de los campos: ***identificacion,nombres, apellidos,correo_electronico,direccion,teléfono_fijo,telefono_celular.***

Para encriptar se debe haber definido una variable de llave que utiliza dos parametros de clave, en este caso se definió **SET @key_str = SHA2('claveDeEncriptamiento' ,512);**

Dicha clave se usa también al momento de desencriptar la misma información

Pre-Insert

```
558    DELIMITER //
559    CREATE OR REPLACE TRIGGER EncriptPaciente
560    Before INSERT
561       ON coronavirus_col.paciente FOR EACH ROW
562    BEGIN
563    SET @key_str = SHA2('claveDeEncriptamiento' ,512);
564
565    set new.identificacion=AES_ENCRYPT(new.identificacion,@key_str);
566    set new.nombres=aes_encrypt(new.nombres,@key_Str);
567    set new.apellidos=aes_encrypt(new.apellidos,@key_Str);
568    set new.direccion=aes_encrypt(new.direccion,@key_str);
569    set new.telefono_fijo=aes_encrypt(new.telefono_fijo,@key_str);
570    set new.correo_electronico=aes_encrypt(new.correo_electronico,@key_str);
571    set new.telefono_celular=aes_encrypt(new.telefono_celular,@key_Str);
572    END //
573    DELIMITER ;
```

Figura 4 Trigger pre-insert para encriptado de información de paciente

Pre-Update

```
DELIMITER //
CREATE OR REPLACE TRIGGER EncriptPacienteUpdate
Before update
   ON coronavirus_col.paciente FOR EACH ROW
BEGIN
SET @key_str = SHA2('claveDeEncriptamiento' ,512);

set new.identificacion=AES_ENCRYPT(new.identificacion,@key_str);
set new.nombres=aes_encrypt(new.nombres,@key_Str);
set new.apellidos=aes_encrypt(new.apellidos,@key_Str);
set new.direccion=aes_encrypt(new.direccion,@key_str);
set new.telefono_fijo=aes_encrypt(new.telefono_fijo,@key_str);
set new.correo_electronico=aes_encrypt(new.correo_electronico,@key_str);
set new.telefono_celular=aes_encrypt(new.telefono_celular,@key_Str);
END //
DELIMITER ;
```

Figura 5Trigger pre-update para encriptado de información de paciente

Los datos Encriptados se visualizan de la siguiente manera:

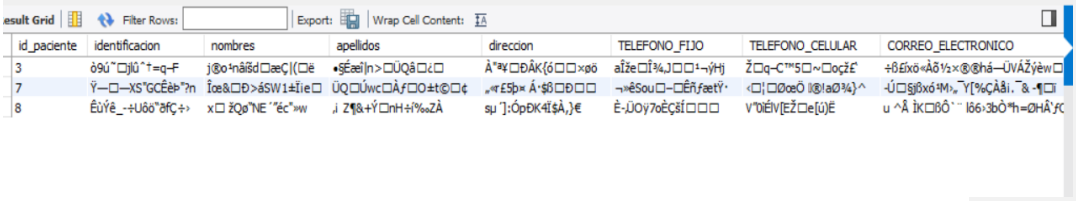

Figura 6 Visualización de información encriptada de paciente

4.4.8 Desencriptado de Información.

Comentado [BEMO9]: ¿y esta sangría?

Para el desencriptado de valores de campos mencionados anteriormente en la tabla: **paciente**, se construyó el siguiente procedimiento almacenado que recibe como parámetro la cadena asociada a llave de encriptado, definida anteriormente para los triggers de Pre-Insert y Pre-Update

```
DELIMITER //
CREATE OR REPLACE PROCEDURE desencriptPaciente (in clave varchar(25))
BEGIN
SET @key_str = SHA2(clave ,512);
select CONVERT(AES_DECRYPT(identificacion,@key_str) USING utf8) as identificacion,
CONVERT(AES_DECRYPT(nombres,@key_str) USING utf8) as nombres,
CONVERT(AES_DECRYPT(apellidos,@key_str) USING utf8) as apellidos,
CONVERT(AES_DECRYPT(direccion,@key_str) USING utf8) as direccion,
CONVERT(AES_DECRYPT(telefono_fijo,@key_str) USING utf8) as telefono_fijo,
CONVERT(AES_DECRYPT(correo_electronico,@key_str) USING utf8) as correo_electronico,
CONVERT(AES_DECRYPT(telefono_celular,@key_str) USING utf8) as telefono_celular
from paciente;
END
// DELIMITER;
```

Figura 7 Procedimiento para desencriptado de información de paciente

Después de hacer una invocación o llamado de dicho procedimiento se obtiene la información en claro de la siguiente manera:

```
712     call desencriptPaciente('claveDeEncriptamiento');
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

identificacion	nombres	apellidos	direccion	telefono_fijo	correo_electronico	telefono_celular
1151950123	OSCAR	FLOREZ	CARRERA 1A 14	4329364	oscardf0812@hotmail.com	3164637291
12233434434	cristian	botina	calle oeste	3328756	cristian@hotmail.com	3128765436
121212121	Luis	Florez	AVENIDA 2 N	3329834	luis@hotmail.com	3187620987

Figura 8 Visualización de información desencriptada de paciente

4.4.9 Auditoría

Con el fin de llevar un registro controlado y monitoreado sobre las tablas transaccionales: ***paciente***, ***estado_paciente*** y ***diagnostico_paciente***, se crearon Triggers que se disparan al momento de realizar movimientos o acciones en datos de tipo INSERT, UPDATE y DELETE sobre la tablas mencionadas, y que dichos movimientos se registren en la tabla: ***auditoriasmovimientos*** de la base de datos: ***db_auditoria_coronavirus_col***, la cual

se creó con el fin de separar la auditoria de la transaccionalidad que presenta la base de datos:***coronavirus_col.***

A continuación se presenta el Trigger que inserta en la tabla ***auditoriasmovimientos*** de la base de datos: ***db_auditoria_coronavirus_col***

```
DELIMITER //
CREATE OR REPLACE TRIGGER audMov_insert_tbldiagnostico_paciente
AFTER INSERT
   ON coronavirus_col.diagnostico_paciente FOR EACH ROW
BEGIN
   DECLARE vUsuario varchar(80);
   DECLARE vBasedatos varchar(100);
   DECLARE vIdafectado int;
   DECLARE vMovimiento text;
   -- seleccionar variables
   SELECT USER() INTO vUsuario;
   SELECT schema() INTO vBasedatos;
   select LAST_INSERT_ID() INTO vIdafectado;
   select concat(NEW.FECHA_DIAGNOSTICO,",", NEW.VALOR,",", NEW.PACIENTE_ID_PACIENTE,",",NEW.DIAGNOSTICO_ID_DIAGNOSTICO) INTO vMovimiento;
-- insertar el registro en la tabla para auditar
   INSERT INTO db_auditoria_coronavirus_col.auditoriaMovimientos
   (tipomovimiento, idafectado, movimiento, usuario, fechaHora, basedatos, objeto)
   VALUES
   ( 'Insert', vIdafectado, vMovimiento, vUsuario, SYSDATE(), vBasedatos, 'coronavirus_col.diagnostico_paciente');
END; //
```

Figura 9 Trigger auditoria para guardar registro tipo insert de diagnóstico - paciente

Al realizar consultas sobre la tabla ***auditoriasmovimientos,*** se podrá conocer que tipo de movimiento(insert,update,delete), Id afectado, movimiento, usuarios que realizó el movimiento, fecha y hora del movimiento, base datos objeto afectado(tabla)

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

id	tipomovimiento	idafectado	movimiento	usuario	fechaHora	basedatos	objeto
4	Insert	1	2020-06-14,1,1,1,3	root@localhost	2020-06-14 19:56:35	coronavirus_col	coronavirus_col.estado_paciente
5	Insert	1	2020-06-01,5.5,3,1	root@localhost	2020-06-14 19:59:50	coronavirus_col	coronavirus_col.diagnostico_paciente
6	update	1	2020-06-01,8.0,3,1	root@localhost	2020-06-14 20:00:59	coronavirus_col	coronavirus_col.diagnostico_paciente
7	update	1	2020-06-14,1,1,2,3	root@localhost	2020-06-14 20:01:30	coronavirus_col	coronavirus_col.estado_paciente
8	update	0	1151950123,OSCAR,FLOREZ,1993-08-12,M,C...	AGAVIRIAF@localhost	2020-06-15 15:00:11	coronavirus_col	coronavirus_col.paciente
9	update	0	1151950123,OSCAR,FLOREZ,1993-08-12,M,C...	root@localhost	2020-06-16 20:04:23	coronavirus_col	coronavirus_col.paciente
17	Insert	0	121212121,Luis,Florez,1989-08-04,M,AVENIDA...	root@localhost	2020-06-18 18:12:01	coronavirus_col	coronavirus_col.paciente
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figura 10 Visualización información de auditoria

Teniendo en cuenta que en la tabla ***paciente***, se encriptaron los campos sensibles de información, los triggers de auditoria para registrar los movimientos insert,update y delete se implementaron de tal forma que desencriptaran los campos que tienen protección(***identificacion,nombres,apellidos,dirección,telefono_fijo,telefono_celular,correo_electronico***) para que dichos valores se inserten desencriptados en la tabla: ***auditoriasmovimientos*** de la base de datos **db_auditoria_coronavirus_col.**

```
DELIMITER //
CREATE OR REPLACE TRIGGER audMov_insert_tblpaciente
AFTER INSERT
   ON coronavirus_col.paciente FOR EACH ROW
BEGIN
   DECLARE vUsuario varchar(80);
   DECLARE vBasedatos varchar(100);
   DECLARE vIdafectado int;
   DECLARE vMovimiento text;
   -- seleccionar variables
   SELECT USER() INTO vUsuario;
   SELECT schema() INTO vBasedatos;
   select LAST_INSERT_ID() INTO vIdafectado;
   select concat(AES_DECRYPT(NEW.IDENTIFICACION,@key_str) ,",", AES_DECRYPT(NEW.NOMBRES,@key_str) ,",",
   AES_DECRYPT(NEW.APELLIDOS,@key_str) ,",",NEW.FECHA_NACIMIENTO,",",NEW.SEXO,",",AES_DECRYPT(NEW.DIRECCION,@key_str) ,",",
   AES_DECRYPT(NEW.TELEFONO_FIJO,@key_str) ,",",AES_DECRYPT(NEW.CORREO_ELECTRONICO,@key_str) ,",",
   NEW.TIPO_IDENTIFICACION_ID_TIPO_IDENTIFICACION,",",NEW.ENTIDAD_SALUD_ID_ENTIDAD_SALUD,",",
   AES_DECRYPT(NEW.TELEFONO_CELULAR,@key_str),",",NEW.ESTRATO_SOCIAL) INTO vMovimiento;
-- insertar el registro en la tabla para auditar
   INSERT INTO db_auditoria_coronavirus_col.auditoriaMovimientos
   (tipomovimiento, idafectado, movimiento, usuario, fechaHora, basedatos, objeto)
   VALUES
   ( 'Insert', vIdafectado, vMovimiento, vUsuario, SYSDATE(), vBasedatos, 'coronavirus_col.paciente');
END; //
DELIMITER ;
```

Figura 11 Trigger auditoría para guardar registro tipo insert de paciente

5. Aplicación de Minería de Datos sobre una fuente externa

El trabajo de minería de datos realizado, se aplicó a una fuente de datos externa de Kaggle (https://www.kaggle.com/roche-data-science-coalition/uncover/data), la cual presenta datos relacionados a diagnósticos realizados a pacientes, cuyo estado de salud se encuentra en observación para determinar si presenta el Covid-19.

El interrogante de este trabajo pretende develar bajo que condiciones dadas por los diagnósticos respiratorios y de orina puede aumentar o disminuir el nivel de confianza en los datos de entrenamiento para determinar si el paciente es positivo o negativo para Covid-19.

Para efectuar dicho análisis se utilizó el software Weka como herramienta de aprendizaje automático de información y a continuación se explican las siguientes etapas o pasos del proceso de minería de datos:

5.1 Selección del conjunto de Datos

Se realizó una identificación de los atributos más relevantes presentes en el conjunto de datos origen, de los 111 atributos que tiene el archivo, se seleccionaron 41 atributos relevantes que describen a grandes rasgos, diagnósticos de influenza, parainfluenza, orina, estado de salud, edad y resultado del test sars_cov2 realizados al paciente.

Fields :

#	Fieldname	Rename to	Length	Precision
1	patient_addmited_to_regular_ward_1_yes_0_no			
2	patient_addmited_to_semi_intensive_unit_1_yes_0_no			
3	patient_addmited_to_intensive_care_unit_1_yes_0_no			
4	respiratory_syncytial_virus			
5	influenza_a			
6	influenza_b			
7	influenza_a_rapid_test			
8	influenza_b_rapid_test			
9	coronavirusnl63			
1..	coronavirus_hku1			
1..	coronavirus229e			
1..	coronavirusoc43			
1..	parainfluenza_1			
1..	parainfluenza_2			
1..	parainfluenza_3			
1..	parainfluenza_4			
1..	rhinovirus_enterovirus			
1..	chlamydophila_pneumoniae			
1..	adenovirus			
2..	inf_a_h1n1_2009			
2..	bordetella_pertussis			
2..	metapneumovirus			
2..	strepto_a			
2..	myeloblasts			
2..	urine_esterase			
2..	urine_aspect			
2..	urine_ph			
2..	urine_hemoglobin			
2..	urine_bile_pigments			
3..	urine_ketone_bodies			
3..	urine_nitrite			
3..	urine_granular_cylinders			

Figura 12 Selección de atributos

5.2 Preprocesamiento

Se identificaron inconsistencias en los datos de la fuente origen, lo que conllevó a utilizar la herramienta Pentaho, para realizar una depuración de registros erróneos y al mismo tiempo seleccionar los atributos relevantes, con el objetivo de tener datos que puedan ser soportados por el aplicativo Weka.

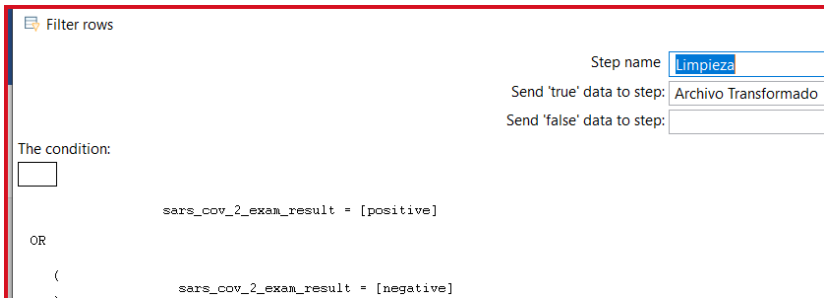

Figura 13 Depuración y limpieza de datos

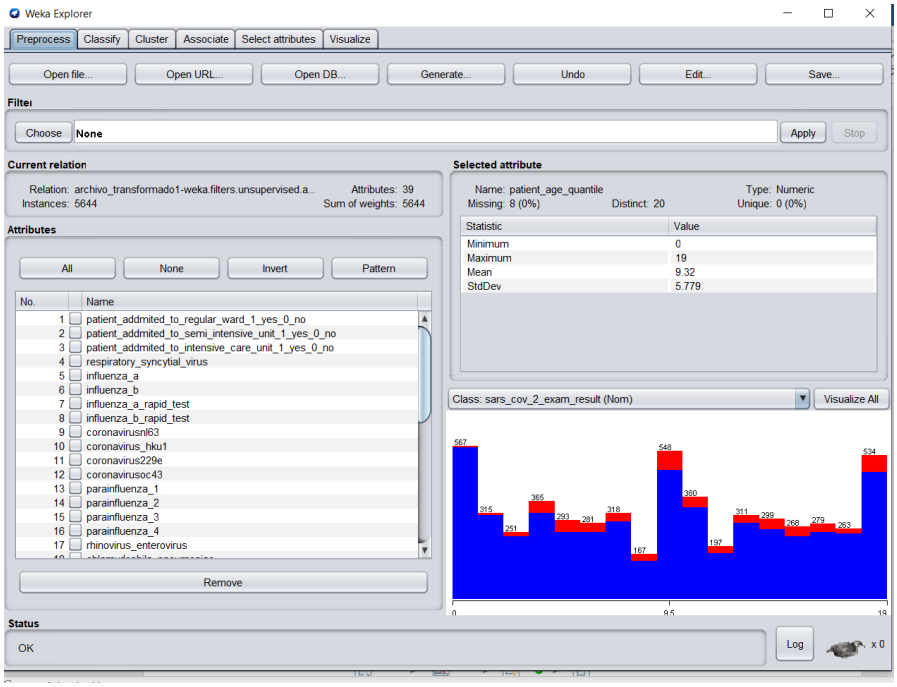

Figura 14 Visualización de atributos en weka

5.3 Transformación del conjunto de datos de Entrada

Se generó un archivo plano de salida, por medio de una ETL, con el fin de tener un recurso de datos preparado para aplicar minería de datos en Weka.

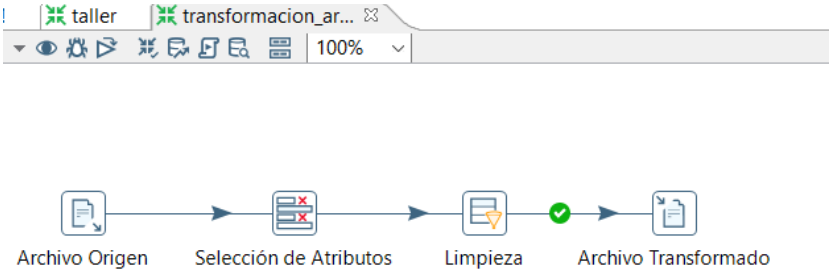

Figura 15 Transformación archivo de entrada

5.4 Seleccionar y aplicar la Técnica de Minería de datos

De acuerdo a la naturaleza de los datos, se aplicó un algoritmo de clasificación en la herramienta Weka, para buscar un modelo que describa el atributo clase como una función de atributos de salida.

En este caso la técnica de minería de datos apropiada fué un árbol de decisiones del algoritmo J48, teniendo en cuenta que lo que se buscaba era llegar a una predicción acertada de los datos siguiendo un modelo lógico basado en reglas, cuya implementación fue la siguiente:

Comentado [BEMO10]: Es importante que si se habla en pasado todo vaya en pasado

Comentado [BEMO11]: fué

Comentado [BEMO12]: buscaba era

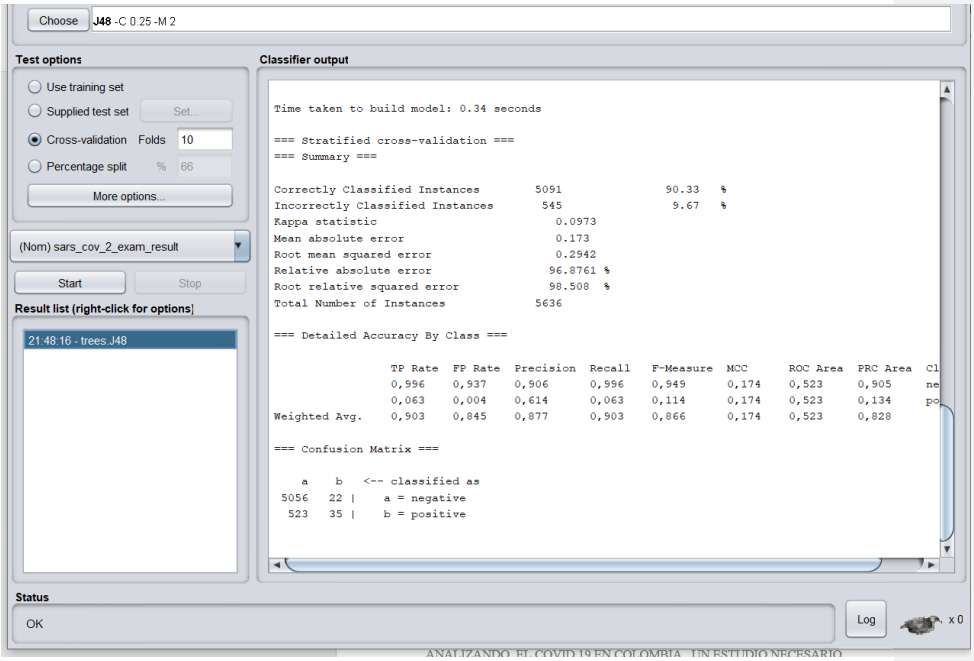

Figura 16 Aplicación algoritmo de clasificación

5.5 Evaluación de Resultados

Correctly Classified Instances	5091	90.33 %
Incorrectly Classified Instances	545	9.67 %
Kappa statistic	0.0973	
Mean absolute error	0.173	
Root mean squared error	0.2942	
Relative absolute error	96.8761 %	
Root relative squared error	98.508 %	
Total Number of Instances	5636	

De acuerdo a los datos presentes en la fuente origen y siguiendo el modelo de predicción obtenido, se identifica que el promedio de edad de los pacientes no supera los 19 años, lo que quiere decir que es una muestra de población joven y que aplicando el algoritmo brinda un modelo acorde a los datos pero al mismo tiempo limita e invita a considerar datos que garanticen más variedad de información para obtener mejores resultados. Por otro lado los casos positivos de Covid 19 no presentan mucha información recurso en los diagnósticos implicados, por ello la mayoría aplican como casos negativos y muestra de ello se puede evidenciar en la matriz de confusión.

5.5.1 Matriz de Confusión

```
   a    b   <-- classified as
5056   22 |    a = negative
 523   35 |    b = positive
```

5.5.2Árbol de Decisión

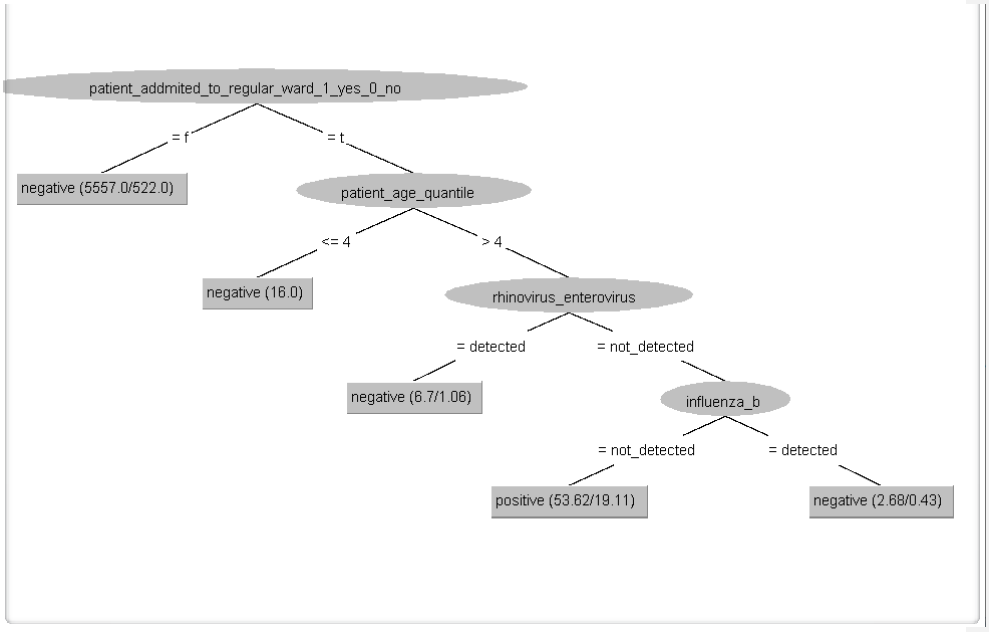

Figura 17 Árbol de decisión algoritmo j48

5 Conclusiones

- Varias de las aplicaciones que manejamos en nuestro diario vivir que requieren confidencialidad necesitan modelos más sofisticados de seguridad: Entidades bancarias, médicas, departamentos gubernamentales, inteligencia militar, etc.
- Aunque la implementación de seguridad más sofisticada no es tarea fácil, debemos hacer el esfuerzo por lograr que nuestros datos estén completamente seguros
- Es importante tener implementación de procesos de auditoría sobre la base de datos.
- En resumen, data mining se presenta como una tecnología innovadora, que ofrece una serie de beneficios: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; en segunda instancia, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- Además, no hay duda de que trabajar con esta tecnología implica cuidar un sin número de detalles debido a que el producto final involucra "toma de decisiones". En este trabajo se ha presentado una alternativa de software de minería de datos Weka, que es una herramienta libre y muy interesante a la hora de aplicar diversas técnicas de minería de datos.
- El éxito de aplicar minería de datos depende del recurso de dato origen.

6 Recomendaciones

- Limitar el acceso a los usuarios y el número de administradores de la base de datos. Respaldar la información de la base de datos eventualmente.
- Implementar métodos de autenticación para no dar libre acceso a la información.
- Implementar seguridad a través de periféricos tanto a nivel de hardware como a nivel de software y conocer de los diferentes ataques que puede sufrir una base de datos y tratar de prevenir estos con anticipación.
- Mantener los servidores de datos bien protegidos en una sala de acceso restringido en donde los administradores puedan acceder mediante acceso remoto.
- Crear diversas normas y procedimientos que determinen exactamente quienes tienen acceso a la información a través de un marco jurídico.
- Aplicar técnicas de minería de datos con los datos que alimenten el modelo diseñado para Colombia

Referencias

«Casos positivos de COVID-19 en Colombia | Datos Abiertos Colombia». *la plataforma de datos abiertos del gobierno colombiano*. Recuperado 3 de junio de 2020a (https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data).

«LEY 1266 DE 2008». Recuperado 16 de junio de 2020b (http://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1676616).

Corso, C. L. 2020 «(PDF) Aplicación de algoritmos de clasificación supervisada usando Weka | Reyes S - Academia.edu». Recuperado 22 de junio de 2020c (https://www.academia.edu/9538896/Aplicaci%C3%B3n_de_algoritmos_de_clasificaci%C3%B3n_supervisada_usando_Weka).

Corso, C. L. 2020 (PDF) COMPARACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA LA PREDICCIÓN DE CASOS DE OBESIDAD INFANTIL». *ResearchGate*. Recuperado 22 de junio de 2020d (https://www.researchgate.net/publication/301567339_COMPARACION_DE_ALGORITMOS_DE_CLASIFICACION_PARA_LA_PREDICCION_DE_CASOS_DE_OBESIDAD_INFANTIL).

Brian, D. A., y R. S. Baric. 2005. «Coronavirus genome structure and replication». Pp. 1–30 en *Coronavirus replication and reverse genetics*. Springer.

López, R. 2018. «MANUAL DE POLÍTICAS DE SEGURIDAD Y PRIVACIDAD DE LA INFORMACIÓN Publicación Institucional Bogotá, D.C., abril de 2018». 79.

Disterer, Georg. 2013. «ISO/IEC 27000, 27001 and 27002 for information security management».

Morales, J. 2020. «OPS/OMS Colombia - Nueva base de datos de la OPS sobre orientación e investigación de la COVID-19 | OPS/OMS». *Pan American Health Organization / World Health Organization*. Recuperado 3 de junio de 2020 (https://www.paho.org/col/index.php?option=com_content&view=article&id=3354:nueva-base-de-datos-de-la-ops-sobre-orientacion-e-investigacion-de-la-covid-19&Itemid=562).

Novel, 2020. «The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China». *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi* 41(2):145.

OMS, Organización Mundial de la Salud. s. f. «Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19)». Recuperado 7 de junio de 2020 (https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses).

Roche Data Science Coalition. 2020 «UNCOVER COVID-19 Challenge». Recuperado 3 de junio de 2020 (https://kaggle.com/roche-data-science-coalition/uncover).

Sierra,B. 2006. Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando eISO27000.es. s. f. «ISO 27000 y el conjunto de estándares de Seguridad de la Información». Recuperado 27 de junio de 2020 (http://www.intedya.com/internacional/757/noticia-iso-27000-y-el-conjunto-de-estandares-de-seguridad-de-la-informacion.html).