UNIVERSIDAD SAN GREGORIO DE PORTOVIEJO

Maestría en Gestión y Analítica de Datos

Gestión de Datos

Paralelo: “A”

ALUMNOS:

López Contreras Jean Carlos

Zamora Macías David Jordan

Quirola Batallas Leonardo Javier

Espinoza Loor Gorki Alexander

PROYECTO INTEGRADOR

DOCENTE: Ing. Adriana Collaguazo Jaramillo, Mg.

PERIODO ACADÉMICO: 2024 – 2025

Portoviejo – Manabí - Ecuador

ÍNDICE

1. Tema:

Análisis de Datos de Salud para la Detección y Predicción de Enfermedades Respiratorias

2. Introducción:

El análisis de datos en el ámbito de la salud representa una oportunidad clave para mejorar la prevención, diagnóstico y tratamiento de enfermedades mediante el uso de tecnologías emergentes. En particular, la detección temprana de enfermedades respiratorias ha cobrado relevancia por su impacto en la calidad de vida de la población y en la carga del sistema sanitario. En este contexto, el presente proyecto, titulado Análisis de Datos de Salud para la Detección y Predicción de Enfermedades Respiratorias, propone una metodología basada en el procesamiento, exploración y modelado de datos clínicos, con el objetivo de identificar patrones relevantes y construir modelos predictivos que apoyen la toma de decisiones médicas.

Para el desarrollo del proyecto se empleó el lenguaje de programación Python, junto con bibliotecas especializadas como pandas para la manipulación de datos, numpy para cálculos numéricos, matplotlib y seaborn para visualización, así como scikit-learn para la implementación de algoritmos de aprendizaje automático. Se inició con una fase de preprocesamiento orientada a la limpieza y transformación de los datos, incluyendo la imputación de valores nulos, la conversión de tipos de datos y la normalización de variables. Posteriormente, se realizó un análisis exploratorio con el propósito de identificar tendencias y relaciones entre las variables más representativas del conjunto de datos clínicos.

Con base en la información procesada, se entrenaron y evaluaron distintos modelos de clasificación, tales como regresión logística, árbol de decisión y perceptrón multicapa, con la finalidad de predecir la presencia de enfermedades respiratorias. Los resultados obtenidos permiten evaluar el desempeño de cada modelo a través de métricas como precisión, recall y F1-score, y brindan evidencia sobre la viabilidad de utilizar enfoques basados en inteligencia artificial para fortalecer los sistemas de salud. Este proyecto constituye un aporte preliminar hacia el desarrollo de herramientas inteligentes que integren datos clínicos con técnicas analíticas avanzadas, orientadas a mejorar la atención médica preventiva.

3. Objetivos:

a. Objetivo General:

Desarrollar un análisis de datos de salud enfocado en enfermedades respiratorias para identificar patrones, factores de riesgo y posibles tendencias mediante técnicas de análisis de datos e inteligencia artificial, contribuyendo a la toma de decisiones en el ámbito de la salud pública.

b. Objetivos Específicos:

- Recolectar y limpiar la data de salud relacionada con enfermedades respiratorias para garantizar su calidad y fiabilidad en el análisis.
- Identificar patrones y correlaciones entre variables como edad, ubicación geográfica, factores ambientales y prevalencia de enfermedades respiratorias.
- Implementar modelos de predicción para anticipar brotes o el aumento de casos, utilizando técnicas de machine learning y análisis estadístico.

4. Alcance:

Este proyecto se enfocará en el análisis de datos de salud relacionados con enfermedades respiratorias, aplicando la metodología KDD (Knowledge Discovery in Databases) para la extracción de conocimiento. Se desarrollarán las siguientes etapas:

- **Selección de datos:** Identificación y recopilación de registros médicos, factores ambientales y datos demográficos relevantes.
- **Preprocesamiento:** Limpieza y transformación de la data para eliminar inconsistencias, valores nulos y sesgos.
- **Transformación:** Aplicación de técnicas de reducción de dimensionalidad, normalización y selección de variables clave para el análisis.
- **Minería de Datos:** Implementación de algoritmos de machine learning y análisis estadístico para la identificación de patrones, correlaciones y factores de riesgo.
- **Interpretación y evaluación:** Generación de reportes, visualizaciones y conclusiones basadas en los resultados obtenidos, proporcionando insights para la toma de decisiones en salud pública.

5. Metodología:

Para el desarrollo de este proyecto, se aplicará la metodología **KDD (Knowledge Discovery in Databases)**, que permite la extracción de conocimiento útil a partir de grandes volúmenes de datos. A continuación, se detallan las fases del proceso:

1. Selección de Datos

Se trabajará con un conjunto de datos relacionados con enfermedades respiratorias, que incluirá variables como:

- Datos demográficos (edad, género, ubicación).
- Historial clínico (diagnósticos previos, factores de riesgo).
- Variables ambientales (calidad del aire, temperatura, humedad).

2. Preprocesamiento de Datos

Para garantizar la calidad de los datos, se realizarán las siguientes tareas:

- Eliminación de datos duplicados o inconsistentes.
- Manejo de valores nulos o atípicos mediante técnicas de imputación.
- Normalización y estandarización de variables para un mejor desempeño en los modelos.

3. Transformación de Datos

Se aplicarán técnicas para mejorar la representatividad de los datos, tales como:

- Selección de características relevantes mediante análisis estadístico.
- Reducción de dimensionalidad utilizando **PCA (Análisis de Componentes Principales)** si es necesario.

4. Minería de Datos

Se implementarán algoritmos de **Machine Learning** para identificar patrones y realizar predicciones:

- **Análisis exploratorio** para encontrar tendencias en la incidencia de enfermedades.
- **Modelos predictivos**, como regresión logística, árboles de decisión y redes neuronales, para prever brotes de enfermedades respiratorias.
- **Evaluación de modelos** utilizando métricas como precisión, recall y F1-score.

5. Interpretación y Evaluación de Resultados

Los hallazgos se presentarán mediante:

- **Visualización de datos** con gráficos e informes interactivos.
- **Interpretación de patrones y tendencias** para la toma de decisiones en salud pública.
- **Recomendaciones** basadas en el análisis de datos y resultados obtenidos.

6. Análisis y Resultados

a) Perfil demográfico y características clínicas confiables

Tras el proceso de limpieza y validación, se dispone de un conjunto de datos robusto con más de 380 mil registros, representativos de la población atendida. Esto permite caracterizar con precisión

factores como edad, sexo, peso, talla y diagnóstico, facilitando una planificación ajustada a la realidad epidemiológica.

b) Presencia de variabilidad clínica sin redundancia

El análisis de correlación muestra que las variables antropométricas (peso, talla, edad) no están altamente correlacionadas entre sí, lo cual indica que cada una aporta información relevante y única. Esta riqueza de datos permite construir perfiles clínicos más precisos para estrategias preventivas y de manejo individualizado.

c) Segmentación poblacional significativa mediante técnicas de clustering

Mediante algoritmos de agrupamiento, se identificaron tres grupos distintos de pacientes con características similares. Esto permite diseñar protocolos diferenciados según perfil, implementar intervenciones específicas (ej. adolescentes con salud mental, adultos mayores con morbilidades respiratorias), y priorizar recursos para atención preventiva o seguimiento proactivo.

d) Valoraciones de tendencia temporal para reforzar vigilancia sanitaria

El análisis por meses permite detectar períodos con mayor carga asistencial. Esta información debe ser utilizada para fortalecer el sistema de vigilancia epidemiológica, distribuir personal y reforzar campañas específicas en los meses de mayor demanda.

e) Identificación de patrones de asociación en salud

A través del análisis de reglas de asociación, se identifican combinaciones frecuentes de características clínicas (como edad, peso y diagnóstico). Estas asociaciones son útiles para desarrollar alertas clínicas automatizadas, tamizajes dirigidos y acciones preventivas personalizadas.

7. Riesgos y Limitaciones

El desarrollo de este proyecto enfrenta ciertos riesgos y limitaciones que pueden afectar la calidad y precisión del análisis de datos sobre enfermedades respiratorias. A continuación, se detallan los principales aspectos a considerar:

a. Riesgos

Calidad de los datos: La presencia de datos incompletos, inconsistentes o sesgados puede afectar la fiabilidad del análisis. Se mitigará mediante técnicas de limpieza y preprocesamiento.

Disponibilidad de la información: Si los datos provienen de fuentes externas (hospitales, registros de salud, sensores ambientales), puede haber restricciones en el acceso o demoras en la actualización de la información.

Privacidad y seguridad de los datos: El manejo de información médica sensible requiere cumplir con normativas de protección de datos, evitando la exposición de información personal sin autorización.

Limitaciones computacionales: El procesamiento de grandes volúmenes de datos y la ejecución de modelos de machine learning pueden requerir recursos computacionales elevados.

Precisión de los modelos predictivos: Los algoritmos utilizados pueden verse afectados por la calidad y cantidad de datos disponibles, lo que podría generar predicciones inexactas.

b. Limitaciones

Datos históricos vs. tiempo real: El análisis se basa en datos históricos, lo que puede limitar la capacidad de detectar cambios súbitos en la incidencia de enfermedades.

Factores externos no considerados: Aunque el análisis incluirá variables ambientales y demográficas, otros factores como políticas de salud, comportamiento social y acceso a tratamientos pueden influir en los resultados y no estar reflejados en los datos.

Dependencia de modelos preentrenados: En caso de utilizar modelos ya existentes, la interpretación de los resultados dependerá de la calidad y sesgo de los datos con los que fueron entrenados previamente.

Generalización de los resultados: Los hallazgos pueden ser representativos solo para la población y el período de tiempo analizados, lo que limita su aplicación a otras regiones o contextos.

c. Medidas de Mitigación

Aplicación de técnicas avanzadas de limpieza y procesamiento de datos.

Uso de modelos explicables y validación con métricas de rendimiento.

Evaluación continua de la calidad de los datos y ajustes en los modelos según sea necesario.

8. Conclusiones

El análisis de datos de salud enfocado en enfermedades respiratorias permitió identificar patrones y tendencias relevantes en la incidencia de estas patologías.

Los modelos de Machine Learning demostraron su potencial para predecir brotes o aumentos de casos, aunque su precisión depende de la calidad y cantidad de datos disponibles.

Se identificaron factores clave como edad, ubicación geográfica y variables ambientales que influyen en la prevalencia de enfermedades respiratorias.

A pesar de las limitaciones en la disponibilidad y calidad de los datos, los hallazgos obtenidos pueden ser utilizados para fortalecer la toma de decisiones en salud pública y mejorar estrategias de prevención.

9. Recomendaciones

Mejorar la calidad de los datos: Se recomienda establecer procesos de recolección más rigurosos para minimizar errores y valores faltantes en los registros de salud.

Incorporar más fuentes de información: Ampliar la base de datos con registros de sensores ambientales, reportes hospitalarios y datos epidemiológicos en tiempo real.

Optimizar los modelos predictivos: Ajustar los algoritmos de machine learning utilizando técnicas de hiperparametrización y validación cruzada para mejorar la precisión de las predicciones.

Implementar una plataforma de monitoreo: Crear un sistema que permita la visualización en tiempo real de los datos y facilite la toma de decisiones por parte de autoridades de salud.

Reforzar la seguridad y privacidad de los datos: Adoptar buenas prácticas en el manejo de información médica, cumpliendo con normativas de protección de datos.