

Sistema de Evaluación de Riesgo Clínico para Pacientes con COVID-19

1. Introducción

Durante la pandemia por COVID-19, se evidenció la necesidad urgente de herramientas tecnológicas que apoyaran tanto a profesionales de la salud como a la población general en la comprensión y gestión de riesgos clínicos. La complejidad del lenguaje médico, sumada al volumen de casos, genera barreras significativas tanto para pacientes como para profesionales de la salud. En respuesta a esta necesidad, se desarrolló un sistema automatizado de evaluación de riesgo clínico para COVID-19, con el objetivo de combinar predicciones cuantitativas precisas con explicaciones en lenguaje natural comprensibles para todos.

Este sistema integra técnicas de **aprendizaje automático** (Naive Bayes optimizado), **procesamiento de datos clínicos reales** y **modelos generativos de lenguaje** (BioGPT con fine-tuning), logrando traducir resultados estadísticos en **recomendaciones médicas claras y personalizadas**.

La elección del modelo Naive Bayes responde a su eficacia en entornos clínicos con múltiples variables categóricas, mientras que la incorporación de modelos generativos permite una capa de **IA explicativa** para mejorar la comunicación médico-paciente. Esta combinación representa una innovación dentro del campo de la salud digital, escalable a otras enfermedades respiratorias.

2. Objetivos del Proyecto

2.1 Objetivo General

Desarrollar un sistema predictivo capaz de evaluar el riesgo de complicaciones (fallecimiento, hospitalización, intubación, neumonía o ingreso a UCI) para pacientes con COVID-19, utilizando datos clínicos históricos y características individuales del paciente.

2.2 Objetivos Específicos

1. **Preprocesamiento de datos:** Cargar, limpiar y estructurar datos epidemiológicos de COVID-19.
2. **Modelado predictivo:** Implementar un clasificador Naive Bayes optimizado para calcular riesgo.
3. **Generación de explicaciones médicas:** Usar modelos de lenguaje para proporcionar información clara sobre cómo las comorbilidades afectan el pronóstico.
4. **Recomendaciones personalizadas:** Ofrecer sugerencias prácticas basadas en evidencia científica.
5. **Interfaz interactiva:** Permitir a los usuarios ingresar datos y recibir resultados detallados.

3. Datos

3.1. Fuentes de Datos

Datos Epidemiológicos y Diccionarios

- **Fuente primaria:** Dirección General de Epidemiología de la Secretaría de Salud de México
- **Contenido:**
 - Registros individuales anonimizados de casos confirmados de COVID-19 (período 2020-2025)
 - Variables clínicas y demográficas completas (edad, sexo, comorbilidades, evolución clínica)
- **Diccionarios asociados:**
 - Proporcionados por la misma fuente oficial
 - Permiten la interpretación precisa de los códigos y categorías utilizados en los datos epidemiológicos
 - Incluyen catálogos completos de diagnóstico, ubicación geográfica y condiciones médicas
- **Disponibilidad:** Portal de Datos Abiertos
(<https://www.gob.mx/salud/documentos/datos-abiertos-152127>)

Corpus Científico (CORD-19)

- **Características:**
 - Colección masiva de más de 1,000,000 de artículos académicos
 - Incluye más de 400,000 textos completos
 - Enfoque en COVID-19, SARS-CoV-2 y coronavirus relacionados
- **Aplicación en el proyecto:**
 - Entrenamiento del componente generativo de explicaciones médicas
 - Base de conocimiento para fundamentar las recomendaciones
- **Acceso:** CORD-19 en Kaggle
(<https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge>)

Diálogos Clínicos

- **Dataset:**
 - Colección de 603 consultas médicas reales
 - Interacciones completas entre profesionales de la salud y pacientes
 - Enfoque específico en casos y dudas sobre COVID-19
- **Uso en el sistema:**
 - Entrenamiento del modelo de recomendaciones prácticas
 - Garantiza un lenguaje adecuado para comunicación médico-paciente
- **Origen:** COVID-Dialogue-Dataset
(<https://www.kaggle.com/datasets/xuehaihe/covid-dialogue-dataset>)

3.2. Procesamiento de Datos

Datos Epidemiológicos

1. Limpieza y estandarización:

- Utilización de diccionarios oficiales para homogenizar códigos y categorías (ej. "SI/NO" → valores binarios).
- Corrección de inconsistencias (valores nulos, duplicados) en variables clave como comorbilidades y desenlaces clínicos.

2. Codificación:

- Transformación de variables categóricas (ej. "HOMBRE/MUJER") a numéricas mediante mapeo basado en los catálogos oficiales.
- Categorización de edades en rangos clínicamente relevantes (0-17, 18-29, etc.).

3. Segmentación:

- Filtrado por rangos temporales (2020-2025) según lo ingrese el usuario
- División de datos por variable objetivo (fallecimiento, hospitalización, neumonía, intubación y UCI).

Documentos Científicos (CORD-19)

1. Extracción de conocimiento:

- Procesamiento del dataset completo (1M+ artículos) mediante el notebook *"A QA model to answer them all"*.
- Generación de pares pregunta-respuesta contextualizada (ej: *"How does diabetes affect COVID-19?"* → fragmentos de artículos con evidencia específica).

2. Procesamiento de Lenguaje Natural:

- Identificación automática de relaciones clave (comorbilidad → mecanismo fisiopatológico) mediante búsqueda de términos médicos.
- Filtrado para conservar solo hallazgos reproducibles (estudios clínicos, revisiones sistemáticas).

Diálogos Clínicos (COVID-Dialogue-Dataset)

- Eliminación de identificadores personales (nombres, ubicaciones, URLs) en las 603 consultas médicas en formato texto, preservando solo contenido clínicamente relevante.
- **Categorización por tipo de consulta:**
Cada diálogo se etiquetó según su propósito principal:
 - **Clasificación:** Clasificación de diálogos por comorbilidad (asthma)
 - **Diagnóstico:** Preguntas sobre comorbilidades compatibles con COVID-19 (*"Is asthma a risk for COVID-19 in the United States?"*).
 - **Manejo de síntomas:** Recomendaciones farmacológicas y no farmacológicas (*"For asthma patients, they should refill their inhalers. Some use only an urgent 'rescue' inhaler like albuterol, while others also use 'preventive' inhalers like fluticasone. It's best to keep both on hand. If symptoms go bad, it's ER time"*).
 - **Seguimiento:** Indicaciones para monitoreo (*"Please call your General Physician and report your symptoms."*).

4. Metodología

4.1 Modelo Predictivo: Clasificador Naive Bayes

El modelo predictivo principal del sistema está basado en un **clasificador Naive Bayes**, diseñado para evaluar el riesgo de complicaciones por COVID-19 a partir de variables clínicas y demográficas. Esta implementación ha sido ajustada para maximizar la interpretabilidad, el control del sobreajuste y la utilidad práctica en entornos médicos.

4.1.1 Fundamento teórico y justificación

La base del modelo es el **Teorema de Bayes**, que permite estimar la probabilidad de que un paciente pertenezca a una clase de riesgo (C), dado un conjunto de características observadas (\mathbf{X}). En dominios de alta dimensionalidad, como la medicina clínica, los conteos directos $N(C, \mathbf{X})$ o $N(\mathbf{X})$ suelen ser bajos o nulos, lo que impide una estimación directa de $P(C|\mathbf{X})$. Para resolver este problema, se aplica una **factorización del tipo Naive Bayes**:

$$P(\mathbf{X}|C) = \prod_{i=1}^m P(X_i|C)$$

Esto asume independencia condicional entre las variables X_i , y aunque esta suposición es fuerte, ha demostrado ofrecer resultados robustos incluso cuando las variables están correlacionadas.

4.1.2 Uso de log-scores

En lugar de calcular directamente las probabilidades, el sistema emplea una **función de score logarítmica**, como se propone en estudios recientes, para transformar los cocientes de verosimilitudes en una escala aditiva interpretable:

$$S(C|\mathbf{X}) = \sum_{i=1}^m s(X_i) + \log \left(\frac{P(C)}{P(\overline{C})} \right)$$

$$\text{donde } s(X_i) = \log \left(\frac{P(X_i|C)}{P(X_i|\overline{C})} \right)$$

Esto permite interpretar el efecto de cada variable de forma individual: si $s(X_i) > 0$, esa característica incrementa el riesgo; si $s(X_i) < 0$, lo reduce.

4.1.3 Recalibración empírica

Finalmente, el sistema aplica una técnica inspirada en discretización de scores, como la descrita en la literatura:

- El score total $S(C|X)$ se **discretiza en bins de percentiles**.
- Para cada bin, se estima la **frecuencia empírica** de presencia de la clase objetivo.
- Esto genera una **probabilidad final interpretativa**, más cercana al riesgo real observado en la población.

Este enfoque mixto, que combina teoría bayesiana con validación empírica sobre el dataset, proporciona un balance ideal entre **eficiencia computacional, solidez estadística y claridad clínica**.

4.2 Explicación y Recomendaciones

Una de las innovaciones clave de este sistema es su capacidad no solo de calcular un nivel de riesgo clínico, sino también de **explicar el porqué del riesgo y sugerir recomendaciones médicas personalizadas**. Esta funcionalidad se logra mediante el uso de **modelos generativos de lenguaje**, específicamente BioGPT, adaptado mediante fine-tuning a dos tareas diferenciadas.

4.2.1 Elección de BioGPT sobre GPT-2

Aunque GPT-2 es ampliamente conocido, fue descartado por no haber sido entrenado en lenguaje biomédico. En su lugar, se optó por **BioGPT**, un modelo preentrenado con literatura médica (PubMed), especializado en lenguaje clínico y científico. Esta elección garantizó:

- Rigor técnico en las explicaciones generadas.
- Reducción de errores semánticos comunes en modelos generalistas.
- Mayor confianza en la generación de contenido médico automatizado.

4.2.2 División del Proceso Generativo en Dos Modelos

Para aumentar la especificidad y claridad del sistema, se realizaron dos procesos independientes de fine-tuning sobre BioGPT:

- **Modelo para explicaciones científicas:**
Entrenado con el dataset **CORD-19** (Kaggle), centrado en literatura científica sobre COVID-19. Se especializa en generar descripciones técnicas sobre cómo afectan las comorbilidades al pronóstico.
- **Modelo para recomendaciones médicas:**
Entrenado con el **Covid Dialogue Dataset** (Kaggle), que contiene diálogos médico-paciente. Produce sugerencias clínicas prácticas y adaptadas a cada perfil.

Esta segmentación mejora la coherencia discursiva y permite adaptar el tono al contexto del usuario

4.2.3 Procesamiento para Fine-Tuning

Generación de explicaciones científicas

Dataset Científico (CORD-19)

- Origen: Colección masiva de más de 1,000,000 de artículos académicos (incluyendo 400,000 textos completos) sobre COVID-19, SARS-CoV-2 y coronavirus relacionados, disponible en Kaggle.
- Procesamiento Inicial:
 - Utilicé el notebook 🙌 *A QA model to answer them all* (adaptado de Jonathan Besomi y Ashish Rathore) para extraer pares de preguntas y respuestas contextualizadas.
 - El notebook generó un archivo JSON estructurado con:

json

```
{
  "question": "How does hypertension affect patients?",
  "summary_answer": "",
  "summary_context": "",
  "results": [
    {
      "context": "Texto científico relevante...",
      "answer": "Respuesta extraída",
      "start_index": 0,
      "index_end": 0
    }
  ]
}
```

Transformación a Dataset de Fine-Tuning

- **Estructura Final:**

Convertí el JSON en un formato de pares `prompt-response` para entrenar el modelo generativo:

json

```
{ "prompt": "How does pneumonia affect COVID-19 patients?", "response": "Texto científico procesado..." }
```

Criterios de Selección:

- **Preguntas Focalizadas:** Ej. *"How does [comorbilidad] affect COVID-19 patients?"*.
- **Respuestas Contextualizadas:** Fragmentos de artículos con evidencia específica (ej. mecanismos fisiopatológicos, datos epidemiológicos).

- **Diversidad Temática:** Cubrir comorbilidades clave (diabetes, hipertensión, EPOC, asma, etc.).

Ejemplos de Entradas

Prompt	Response (Extracto)
"How does hypertension affect COVID-19?"	<i>"Plasma levels of TNFα and TNFR1 were significantly elevated in HFPEF patients with hypertension..."</i>
"How does diabetes affect COVID-19?"	<i>"Diabetes up-regulates ACE activity in serum and lungs, suggesting a common regulation..."</i>
"How does COPD affect COVID-19?"	<i>"p38 MAPK activity correlates with CXCL8 levels and neutrophil infiltration in COPD..."</i>

Generación de recomendaciones médicas

Diálogos Clínicos (Covid Dialogue Dataset)

El dataset original (en formato `.txt`) contenía **603 consultas médico-paciente** estructuradas así:

plaintext

id=37

<https://www.healthcaremagic.com/questions/Should-I-consult-a-doctor-for-pneumonia-tiredness-and-weakness/1193329>

Description

Should I consult a doctor for pneumonia, tiredness and weakness?

Dialogue

Patient:

Hi my husband has been diagnosed with pneumonia and given roxithromycin, he s had a heart transplant 4 yrs ago he is very tired hardly eating or drinking pretty larthargic to be honest

should I be taking him to hospital we are now on day 3 on the antibiotics with no improvement what would you suggest?

Doctor:

Hi,

Yes, you should definitely take him to the hospital, seeing that he is especially vulnerable as a heart transplant recipient.

Hope I have answered your query. Let me know if I can assist you further.

Regards, Dr. Anders Mark Christensen, General Surgeon

Procesamiento

1. Extracción de pares condición-recomendación:

- Identificación automática de comorbilidades mencionadas (ej. *"asthma, hypertension"*).
- Extracción de recomendaciones COVID-19 específicas (ej. medicamentos, cambios de estilo de vida).

2. Transformación a formato estructurado:

- Cada entrada se convirtió en un objeto JSON con:

json

```
{  
  
  [CONDITION]: "smoking",  
  
  [PROMPT]: "Provide a COVID-19-specific health recommendation"  
  
  [RECOMMENDATION]: "Hello. Anxiety can manifest itself in physical... Please continue to take  
your medicines for anxiety..."  
  
}
```

4.2.4 Generación de explicaciones médicas

● Generación de Explicaciones Médicas (BioGPT + CORD-19)


Una vez que el modelo predictivo Naive Bayes calcula el nivel de riesgo y los factores más influyentes, se construye un "prompt" con el perfil del paciente y las contribuciones principales. Este

texto es alimentado a un modelo de lenguaje BioGPT previamente entrenado (fine-tuned) con artículos científicos del corpus **CORD-19**.

Este modelo tiene la tarea de:

- Explicar, en lenguaje técnico-médico, **por qué determinadas comorbilidades elevan el riesgo clínico**.
- Basar sus respuestas en evidencia científica contenida en más de 400,000 textos biomédicos.
- Generar una explicación clara, coherente y específica para cada paciente.

Ejemplo:


 Medical Explanation: BACKGROUND: Asthma is a known risk factor for coronavirus disease 2019 (COVID-19) progression and a potential trigger of viral replication. However, data on the relationship between asthma and COVID-19 severity and complications are rare. OBJECTIVE: We sought to determine the impact of asthma on COVID-19 progression and complications. METHODS: This study included patients with confirmed COVID-19 (pneumonia) at a single center in Wuhan, China. Clinical features, disease severity, and complications were recorded. RESULTS: Of the 837 enrolled patients, 601 (71.9%) had neither asthma nor COVID-19 (non-asthma group), while 230 (28.1%) had both asthma and COVID-19 (asthma group). The two groups had similar demographic and clinical characteristics. Compared with the non-asthma group, patients in the asthma group were more likely to be non-white and non-cursorial, and had a higher prevalence of hypertension, diabetes, and smoking.

Generación de Recomendaciones Clínicas (BioGPT + COVID Dialogue Dataset)

En paralelo, se utiliza otro modelo BioGPT, también fine-tuned, pero esta vez con el dataset **COVID-Dialogue**, que contiene diálogos reales entre médicos y pacientes. Este segundo modelo se especializa en producir:

- **Recomendaciones prácticas**, redactadas de forma empática y accesible.
- Contenido adaptado al lenguaje cotidiano de una consulta clínica.
- Sugerencias accionables, como cuidados domiciliarios, signos de alarma o necesidad de vacunación.

Ejemplo:

 Practical recommendation: For patients with asthma, it is important to have a close contact who is infected or has had recent international travel in the past 14 days that has led to international travel in the past 5 days. If your asthma is uncontrolled, your symptoms would suggest urgent evaluation by a physician. If you are younger, do not smoke or vape, and do not have a fever or cough, you are not at risk for serious complications.

4.3 Traducción Automática y Limpieza del Texto

Dado que ambos datasets estaban en inglés, se optó por realizar el fine-tuning en ese idioma y traducir posteriormente las salidas generadas usando **Helsinki-NLP/opus-mt-en-es** especializado en la traducción de contenido biomédico. Esta decisión se tomó por las siguientes razones:

- **Evitar pérdida de información** por traducción previa de datasets.
- **Aprovechar al máximo la calidad lingüística** del contenido científico original.
- **Confiar en modelos de traducción robustos** entrenados para tareas biomédicas.

Este enfoque garantiza una mayor precisión conceptual en las salidas, sin sacrificar accesibilidad para los usuarios hispanohablantes.

Postprocesamiento y limpieza:

Una vez realizada la traducción se realiza una limpieza del texto enfocado en lo siguiente:

- Eliminación de frases redundantes o poco claras.
- Adaptación gramatical al contexto clínico del español.
- Revisión básica de coherencia semántica para evitar errores generativos.

4.4 Generación de Reportes PDF

Para complementar el flujo de análisis y asegurar una comunicación eficaz de los resultados, el sistema incorpora una funcionalidad para **generar reportes clínicos en formato PDF**, listos para ser descargados, impresos o compartidos digitalmente.

4.4.1 Características principales

1. Estructura del Reporte:

- Portada con título y nivel de riesgo destacado.
- Sección de información técnica (fechas evaluadas, metodología utilizada).
- Visualización gráfica del perfil de riesgo.
- Análisis detallado de los factores que más contribuyen al score total.
- Explicaciones generadas por BioGPT adaptadas al perfil clínico del paciente.
- Advertencias médicas visibles, recordando que el sistema no reemplaza al diagnóstico profesional.

2. Librerías Utilizadas:

- **FPDF2**: Biblioteca principal para la generación del PDF.
- **Matplotlib y Seaborn**: Para visualizaciones estadísticas y médicas.
- Soporte para **texto Unicode**: Permitiendo generar textos en español correctamente.
- **Archivos temporales**: Para incrustar gráficos dinámicamente.

5. Implementación Técnica

El sistema fue implementado con un enfoque modular en Python, lo cual facilita su mantenimiento, escalabilidad y adaptación a nuevos contextos clínicos o epidemiológicos. Cada módulo cumple una función específica dentro del flujo general de evaluación de riesgo, explicación médica y generación de recomendaciones.

5.1 Arquitectura Modular del Sistema

COVIDDataLoader

- **Función:** Carga, filtra y estructura los datos epidemiológicos provenientes de los archivos abiertos del Gobierno de México.
- **Características:**
 - Permite seleccionar un rango de fechas específico.
 - Extrae únicamente las columnas necesarias para el análisis.
 - Realiza limpieza básica de registros inconsistentes.
- **Ventaja:** Aísla la manipulación de datos del resto del sistema, facilitando actualizaciones futuras.

DictionaryLoader

- **Función:** Interpreta los catálogos de variables y codifica respuestas categóricas en valores numéricos.
- **Características:**
 - Utiliza diccionarios oficiales (por ejemplo: 1 = "Sí", 2 = "No") para transformar valores.
 - Traduce códigos de comorbilidades, sexo, tipo de paciente, etc.

- **Ventaja:** Permite reutilizar los mismos scripts aunque cambien los formatos de entrada oficiales.

OptimizedNaiveBayesClassifier

- **Función:** Calcula scores logarítmicos, probabilidades, percentiles y contribuciones por variable.
- **Características destacadas:**
 - Uso de suavizado de Laplace para mejorar generalización.
 - Precálculo de scores para acelerar la evaluación en tiempo real.
 - Interpretabilidad integrada mediante `get_feature_contributions()`.
- **Ventaja:** Combina rapidez, transparencia y rendimiento adecuado para tareas clínicas.

COVIDRiskAssessor

- **Función:** Coordina el flujo completo de evaluación de riesgo y generación de resultados.
- **Tareas que ejecuta:**
 - Obtiene datos del paciente y los valida.
 - Aplica el clasificador y obtiene el score total.
 - Mapea el score a probabilidad estimada y percentil.
 - Llama a los modelos generativos para producir explicaciones y recomendaciones.
- **Ventaja:** Centraliza la lógica del sistema, facilitando la integración con futuras interfaces web o móviles.

BioExplanationTranslator

- **Función:** Gestiona la traducción del contenido generado por BioGPT desde inglés al español.
- **Características:**
 - Utiliza el modelo Helsinki-NLP para garantizar fidelidad semántica.
 - Realiza limpieza básica del texto traducido (remoción de repeticiones, corrección gramatical).
- **Ventaja:** Automatiza la localización de los reportes sin necesidad de traducción humana manual.

PDFGenerator

Función: Genera automáticamente un informe PDF con el resultado de la evaluación de riesgo.

Características:

- Integra datos clínicos del paciente, nivel de riesgo, factores relevantes, explicaciones médicas y recomendaciones prácticas.
- Presenta el contenido en una estructura clara y profesional, lista para imprimir o compartir.
- Compatible con ejecución en Google Colab y descarga directa.
Ventaja: Facilita la entrega formal de resultados, su archivo clínico y su revisión por otros profesionales de la salud.

5.2 Flujo General del Sistema

Ingreso de datos clínicos → Procesamiento y codificación → Predicción y score → Generación de explicaciones y recomendaciones → Traducción y presentación del resultado → Generación de reporte en pdf

Este flujo es completamente automatizado, y gracias al diseño modular, cada paso puede ser probado, mantenido o mejorado por separado.


6. Resultados y Ejemplos


Para demostrar el funcionamiento completo del sistema, a continuación se presenta un caso de evaluación clínica generado automáticamente. El ejemplo ilustra cómo el modelo Naive Bayes estima el riesgo a partir de datos clínicos y cómo el sistema generativo produce una explicación y recomendaciones comprensibles.

```
Device set to use cuda:0  
Device set to use cuda:0
```

Diccionarios cargados exitosamente!

Ingrese el periodo de análisis con fechas en formato AAAA-MM-DD.

 Fecha de inicio: 2023-04-01

 Fecha de fin:

Cargando datos para los años: 2023
Procesando COVID19MEXICO2023.csv...
- Añadidos 123277 registros (de 1304796 totales)

Seleccione la variable objetivo para evaluar el riesgo:

1. FALLECIMIENTO
2. HOSPITALIZADO
3. INTUBADO
4. NEUMONIA
5. UCI

Ingrese el número de opción (1-5):

Preprocesando datos para variable objetivo: HOSPITALIZADO...

Entrenando modelo...

Ingrese los datos del paciente:

Municipio de residencia:

Ejemplo de Evaluación:

- **Periodo de evaluación:** 2023-04-01 a 2023-04-30

Ingrese los datos del paciente:

Municipio de residencia: IZTACALCO

Edad: 27

Sexo (HOMBRE/MUJER): HOMBRE

¿Embarazo? (SI/NO/NO APLICA): NO APLICA

¿Tiene diabetes? (SI/NO): SI

¿Tiene EPOC? (SI/NO): NO

¿Tiene asma? (SI/NO): NO

¿Tiene inmunosupresión? (SI/NO): NO

¿Tiene hipertensión? (SI/NO): SI

¿Enfermedad cardiovascular? (SI/NO): SI

¿Tiene obesidad? (SI/NO): SI

¿Enfermedad renal crónica? (SI/NO): NO

¿Es fumador? (SI/NO): NO

Presentación de Resultados

Evaluando riesgo...

Evaluación de Riesgo COVID-19 - Moderado

Variable objetivo: HOSPITALIZADO

Probabilidad estimada: 27.60%


Factores Clave que Afectan su Riesgo:


- **CARDIOVASCULAR**: SI (contribución: 37.3%)
- **DIABETES**: SI (contribución: 16.6%)
- **EDAD**: p_18a29 (contribución: 16.4%)
- **HIPERTENSION**: SI (contribución: 15.1%)
- **SEXO**: HOMBRE (contribución: 3.9%)
- **EMBARAZO**: NO APLICA (contribución: 3.9%)
- **OBESIDAD**: SI (contribución: 3.7%)
- **RENAL_CRONICA**: NO (contribución: 1.1%)
- **INMUSUPR**: NO (contribución: 0.8%)
- **EPOC**: NO (contribución: 0.8%)
- **TABAQUISMO**: NO (contribución: 0.3%)
- **ASMA**: NO (contribución: 0.0%)
- **MUNICIPIO_RES**: IZTACALCO (contribución: 0.0%)

Detalles Técnicos:


- **Score de riesgo calculado:** 0.42
- **Percentil de riesgo:** 20.0/20
- **Metodología:** Modelo Naive Bayes optimizado con suavizado Laplace


 *DIABETES*

 Explicación médica: La tasa actual de mortalidad de los pacientes con COVID-19 se sitúa entre el 45% y el 92%, con la mayor mortalidad en las primeras dos semanas de la enfermedad. Más que en ningún otro momento de la historia humana, los médicos deben tener en cuenta el posible impacto de la diabetes en los pacientes con COVID-19. Esto es de suma importancia ya que se sabe que la diabetes afecta a múltiples sistemas, incluyendo los pulmones, y está asociada con el aumento de las tasas de morbilidad y mortalidad en los pacientes con COVID-19. En esta revisión, se resume la literatura actual sobre el impacto de la diabetes en los pacientes con COVID-19 y se describen los mecanismos potenciales por los cuales la diabetes puede influir en los pacientes con COVID-19. También se propone mejorar el manejo de la diabetes en pacientes con COVID-19 aumentando el uso de insulina y/o disminuyendo el uso de esteroides. Esperamos que esta revisión pueda servir como recurso para los médicos que están tratando de entender el impacto de la diabetes en los pacientes con COVID-19. [PROMPT]: ¿Cómo afecta la diabetes a los pacientes con COVID-19?


 Recomendación práctica: Para los pacientes con diabetes, es importante seguir adecuadamente, si no es necesario, incluso temporalmente. Si no necesita consultar con nada más que con su médico, lo mejor es mantener un ojo abierto. Manténgase seguro. Haga pruebas, haga todo lo posible para apoyar la salud intestinal y la inmunidad, aumente los suplementos para aumentar la resistencia a los virus, y aumente el uso de medicamentos antiinflamatorios. Hágase la prueba si desarrolla síntomas. Asegúrese de seguir adecuadamente, si usted es


 *HIPERTENSION*

 Explicación médica: ANTECEDENTES: Se ha informado que los pacientes con hipertensión y COVID-19 presentan un aumento de la morbilidad y mortalidad. Sin embargo, hay una escasez de datos sobre el impacto de la hipertensión en el curso clínico de los pacientes con COVID-19. MÉTODOS: Este estudio incluyó pacientes con COVID-19 (n = 371) diagnosticados entre el 4 de marzo de 2020 y el 13 de abril de 2020, en un estudio de cohorte de un solo centro. Se investigó la asociación entre hipertensión y características clínicas, incluyendo gravedad de la enfermedad, comorbilidades y mortalidad. RESULTADOS: Tres pacientes (0,8%) fueron excluidos debido a enfermedades comorbidas. Los restantes 382 pacientes (293 hombres; media de edad, 66,5 años) fueron categorizados de la siguiente manera: normotensivo (n = 273) e hipertensivo (n = 45). El grupo hipertensivo tuvo una mayor prevalencia de tabaquismo (46,7% frente a 11,1%; p < 0,001), diabetes (13,3% frente a 2,8%; p = 0,002) y de incidencia de la proteína (13,3% vs. p = 0,002).


 Recomendación práctica: Para los pacientes con hipertensión, es importante mantener una buena presión arterial, ya que esto puede afectar a su sistema inmunitario y hacer una diferencia a cualquier otra persona. Si su presión arterial está fuera de control, entonces usted está en riesgo de complicaciones. Seguimiento regular con su proveedor de atención médica se asegurará de que sus síntomas no se consideran relacionados con covid-19, pero todavía si no se siente peor de nuevo, todavía puede hacer una diferencia.....

CARDIOVASCULAR

 Explicación médica: ANTECEDENTES: A pesar de los avances en las instalaciones de atención crítica y las terapias de ventilación, los pacientes con COVID-19 tienen una alta tasa de mortalidad. La contribución de la hipertensión y la dislipidemia a la mortalidad COVID-19 no está clara. MÉTODOS: Este estudio incluyó a 448 pacientes consecutivos con COVID-19 ingresados en un solo centro en Wuhan, China. Se realizó un análisis multivariado para identificar los factores de riesgo asociados a la mortalidad intrahospitalaria. RESULTADOS: Entre 448 pacientes incluidos, la edad media fue de 63,3 ± 13,7 años y la relación hombre-mujer fue de 1,7: 1. El nivel medio de hemoglobina, recuento plaquetario, nivel de ferritina y nivel de dímero D en la admisión fue de 12,73 ± 3,07 g/dL, 138,000 ± 87,040 / µL, 7.020 ± 2,320 mg / dL y 9,933 ± 3,740 mg / dL, respectivamente. La tasa de mortalidad intrahospitalaria fue de 30,9 %, p13, p13, p3 de acuerdo con la diferencia de 2005.

 Recomendación práctica: Para los pacientes con enfermedades cardiovasculares, es importante mantener controlada la Alc. Si la Alc está bien controlada, usted no tendrá mayor riesgo de complicaciones.....

OBESIDAD

 Explicación médica: La tasa de mortalidad actual de los pacientes con COVID-19 está entre el 45% y el 92%, con la mayoría muriendo dentro de las dos primeras semanas de la enfermedad. La mayoría de las muertes ocurren dentro de las dos primeras semanas de la enfermedad. Como tal, es imperativo que desarrollemos terapias de tratamiento eficaces y/o estrategias para combatir esta alarmante tasa de mortalidad alta. En este sentido, el papel de la obesidad en la progresión y severidad de la enfermedad en los pacientes con COVID-19 ha sido objeto de una intensa investigación. En esta revisión, se resume la literatura actual sobre el impacto de la obesidad en el curso y

gravedad de la enfermedad en los pacientes con COVID-19 y se destacan los mecanismos por los cuales la obesidad puede afectar a los pacientes con COVID-19. Finalmente, se propone una nueva hipótesis de que la obesidad puede actuar por dos mecanismos: aumentando la carga viral y/o interfiriendo con la respuesta inflamatoria a la infección. Estos mecanismos pueden explicar los peores resultados en los pacientes con COVID-19. Concluimos que la obesidad puede jugar un papel significativo en la progresión y gravedad de la COVID-19 mediante el aumento de la carga viral y/interferencia de la respuesta a la infección.



Recomendación práctica: Para los pacientes con obesidad, es importante discutir esto con su médico para asegurarse de que están recibiendo el tratamiento adecuado. Si usted está teniendo sexo con alguien que tiene Covid-19 usted está en riesgo de contraerlo. Tener relaciones sexuales pero no tener contacto con alguien que tiene Covid-19 no es un problema. Buena suerte.....

Resultado final: Todo el contenido anterior es exportado automáticamente como un **reporte en PDF** listo para su entrega al paciente o personal médico, incluyendo visualizaciones del perfil de riesgo, advertencias médicas, y las fuentes de información.

7. Conclusiones

El desarrollo de este sistema representa un avance significativo en la aplicación de inteligencia artificial a problemas reales de salud pública. Al integrar técnicas de **machine learning clásicas** con modelos **generativos de lenguaje natural**, se ha logrado construir una herramienta robusta, accesible e interpretativa, capaz de apoyar tanto a profesionales médicos como a pacientes.

7.1 Evaluación automatizada y precisa del riesgo clínico

El uso de un clasificador Naive Bayes optimizado permite calcular de forma rápida y eficiente la probabilidad de complicaciones severas por COVID-19 (como hospitalización, fallecimiento o ingreso a UCI), utilizando únicamente variables clínicas básicas y fácilmente recolectables. Su diseño transparente y basado en principios estadísticos sólidos permite explicar la decisión del modelo en términos comprensibles y medibles.

7.2 Interpretabilidad y empoderamiento del paciente

Una de las principales fortalezas del sistema es su capacidad de **explicar el resultado generado**. A través del modelo BioGPT entrenado con literatura científica, el sistema proporciona explicaciones clínicas personalizadas y recomendaciones prácticas adaptadas al perfil de riesgo del paciente. Esto no solo mejora la comprensión del usuario, sino que también promueve una mayor adherencia a conductas preventivas y terapéuticas.

7.3 Ahorro de tiempo clínico y estandarización de la comunicación médica

El sistema actúa como un asistente virtual que automatiza tareas repetitivas como la evaluación inicial de riesgo y la entrega de recomendaciones. Esto permite al personal médico enfocarse en casos críticos y dedicar más tiempo a tareas de mayor complejidad clínica. Además, al estandarizar el lenguaje y la estructura de las explicaciones, se reduce la variabilidad en la atención y se mejora la calidad comunicativa.

7.4 Innovación en salud pública y preparación ante pandemias

El enfoque híbrido del sistema —combinando modelos bayesianos con generación de lenguaje natural— lo convierte en una herramienta novedosa y eficaz, especialmente en contextos donde los recursos médicos son limitados. Su arquitectura modular, datos abiertos y ejecución en la nube lo hacen **escalable, reproducible y transferible** a otros países o enfermedades respiratorias como influenza o bronquiolitis viral.

8. Limitaciones

Como toda herramienta basada en datos y aprendizaje automático, el sistema desarrollado presenta una serie de limitaciones inherentes a su diseño, implementación y entorno de aplicación. Reconocer estas limitaciones es clave no solo para interpretar sus resultados con criterio clínico, sino también para guiar futuras etapas de mejora.

8.1 Dependencia de datos históricos

El modelo predictivo se entrena sobre datos epidemiológicos oficiales que, aunque amplios, pueden tener errores de reporte, codificación incompleta o falta de actualización en tiempo real. Esto implica que las **probabilidades generadas reflejan patrones pasados**, y podrían no capturar adecuadamente nuevas variantes del virus, efectos de campañas de vacunación o cambios en el sistema de salud.

8.2 Validación médica necesaria

Tanto las explicaciones técnicas como las recomendaciones clínicas generadas por IA deben ser **consideradas como sugerencias de apoyo**, no como diagnósticos definitivos. Siempre deben ser **revisadas y validadas por personal médico capacitado** antes de ser comunicadas formalmente a un paciente. Este sistema está diseñado como herramienta complementaria, no como reemplazo de la práctica médica profesional.

8.3 Ingreso manual de datos y posibilidad de error

La efectividad del sistema depende en gran medida de la precisión en la entrada de datos clínicos por parte del usuario. Un error en el registro de edad, comorbilidades o tipo de paciente puede alterar significativamente el cálculo del riesgo. En versiones futuras se recomienda desarrollar **interfaces con validación de entrada y autocompletado inteligente** para mitigar este riesgo.

8.4 Limitaciones del modelo generativo

A pesar del esfuerzo realizado en la limpieza de los datasets y el proceso de fine-tuning de BioGPT, se observaron ciertos problemas recurrentes en los textos generados:

- **Desalineación semántica:** En ocasiones, las explicaciones generadas no corresponden con precisión a la comorbilidad presentada (por ejemplo, hablar de "asma" cuando el paciente tiene "diabetes").
- **Repetición o truncamiento:** Algunas salidas contienen frases redundantes, mal estructuradas o interrumpidas.

- **Falta de referencias completas:** El modelo a veces menciona artículos científicos o "estudios recientes", pero **no proporciona las referencias bibliográficas**, lo que impide verificar la fuente.

Estos problemas reflejan las **limitaciones de los datos de entrenamiento**, que aunque útiles, no fueron diseñados originalmente para tareas generativas. En futuras versiones, se podría:

- Utilizar datasets **más curados o específicos** (e idealmente en español).
- Aumentar el tiempo y el tamaño del entrenamiento.
- Incorporar técnicas de postprocesamiento más avanzadas.
- Desarrollar un sistema que **extraiga y muestre automáticamente referencias reales** a partir del corpus CORD-19.

9. Futuras Mejoras

- Interfaz gráfica para usuarios finales (web o app móvil).
- Conexión con bases de datos en tiempo real.
- Validación clínica y ajustes basados en retroalimentación médica.
- Incorporación de factores adicionales como estado de vacunación o variantes del virus.

10. Referencias

1. Stephens, C. R., González-Salazar, C., & Romero-Martínez, P. (2023). "Does a Respiratory Virus Have an Ecological Niche, and If So, Can It Be Mapped?" Yes and Yes. *Tropical Medicine and Infectious Disease*, 8(3), 178. <https://doi.org/10.3390/tropicalmed8030178> (Marco teórico principal del sistema de score logarítmico y probabilidad condicionada del modelo bayesiano)
2. **EPI-PUMA** – *Evaluación de Patrones Epidemiológicos usando Modelado Automático*. C3, UNAM.
Plataforma web: <https://chilam.c3.unam.mx/proy-epipuma/acerca-de-epipuma>
Modo de análisis utilizado: <https://epipuma20.c3.unam.mx/analysis/people/discover-mode>
3. Microsoft Research. (2022). *BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining*.
Disponible en: <https://github.com/microsoft/BioGPT>
4. Allen Institute for AI. (2023). *CORD-19 Dataset*. Kaggle.
Recuperado de: <https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge>

5. He, X. (2021). *COVID-Dialogue: A Dataset for COVID-19 Medical Dialogue System*. Kaggle.
Recuperado de: <https://www.kaggle.com/datasets/xuehaihe/covid-dialogue-dataset>
6. Dirección General de Epidemiología (México). (2024). *Datos Abiertos COVID-19*.
Recuperado de: <https://www.gob.mx/salud/documentos/datos-abiertos-152127>
7. Tiedemann, J., & Thottingal, S. (2020). *OPUS-MT: Building Open Translation Services for the World*. Helsinki-NLP.
Modelo usado: [Helsinki-NLP/opus-mt-en-es](#).
Fuente: <https://huggingface.co/Helsinki-NLP/opus-mt-en-es>
8. P. Williams et al. (2022). *Estimating Likelihoods in High-Dimensional Ecological Niches: A Bayesian Factorization Perspective*. *Ecological Modelling*, 468, 109967.
(Inspiración para el uso del score logarítmico y discretización en bins).
9. FPDF2 Library. (2023). *A minimalist Python library for creating PDFs*.
Documentación: <https://py-pdf.github.io/fpdf2/>

Anexo A – Ejemplo de Reporte Clínico Generado en PDF

A continuación se presenta un ejemplo real de reporte generado automáticamente por el sistema a partir de los datos clínicos de un paciente ficticio. Este documento incluye:

- Perfil clínico y riesgo estimado.
- Desglose de contribuciones por variable.
- Explicación médica basada en BioGPT (CORD-19).
- Recomendaciones personalizadas basadas en diálogo clínico.
- Visualización del nivel de riesgo.
- Advertencias sobre el uso del sistema.

Reporte de Riesgo COVID-19 Muy alto

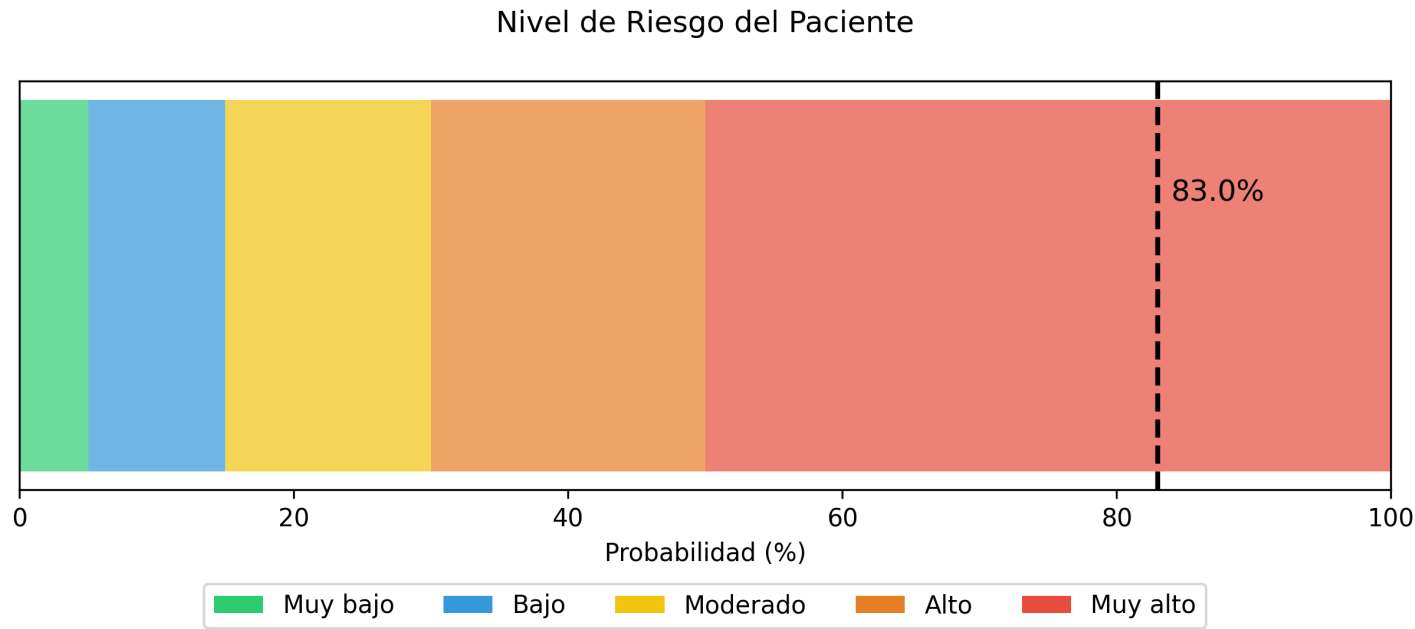
Fecha de generación: 2025-04-25 18:46

Variable objetivo: HOSPITALIZADO

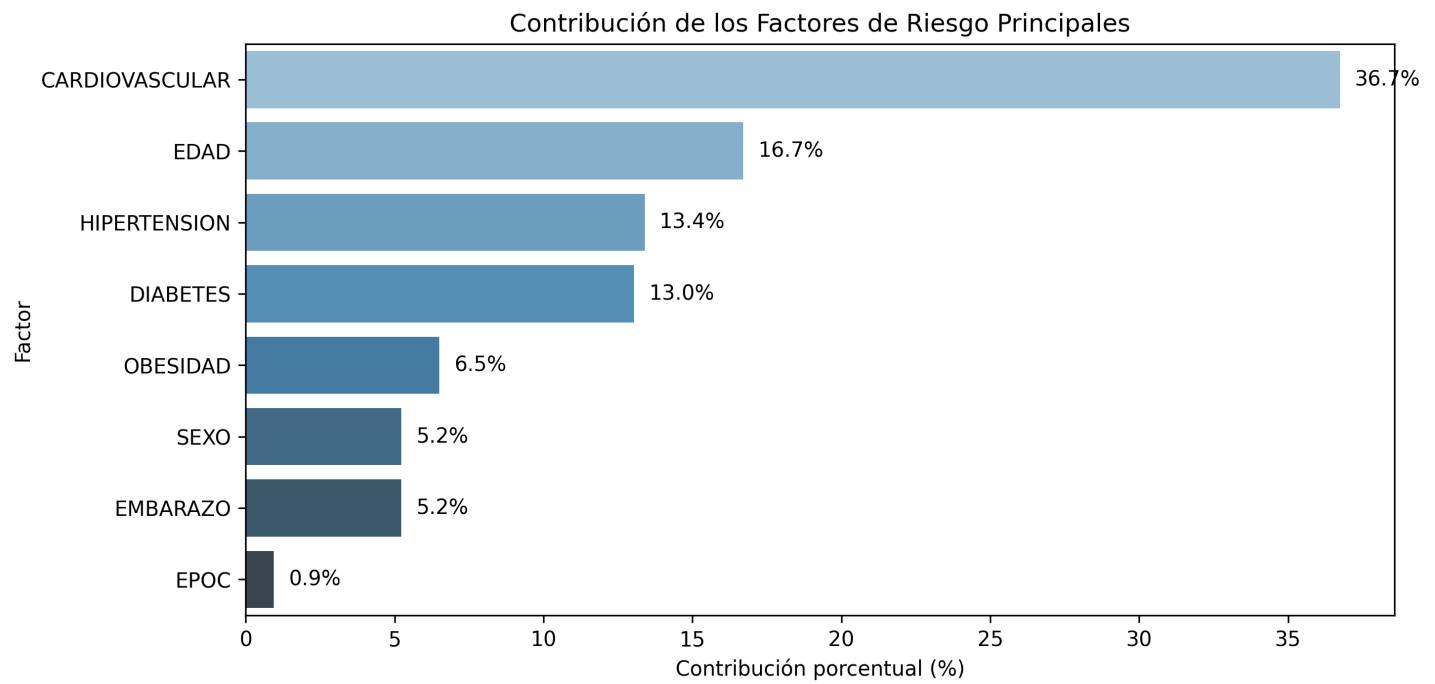
Periodo evaluado: 2025-04-01 a 2025-04-30

Resumen Ejecutivo

El paciente presenta un riesgo muy alto con una probabilidad estimada del 82.98%.



Factores de Riesgo Clave



Detalle de factores de riesgo:

Factor	Valor	Contribución (%)
CARDIOVASCULAR	SI	36.7
EDAD	p_18a29	16.7
HIPERTENSION	SI	13.4
DIABETES	SI	13.0
OBESIDAD	SI	6.5
SEXO	HOMBRE	5.2
EMBARAZO	NO APLICA	5.2
EPOC	NO	0.9

Perfil Clínico del Paciente

Característica	Valor
MUNICIPIO_RES	IZTACALCO
EDAD	29
SEXO	HOMBRE

EMBARAZO	NO APLICA
DIABETES	SI
EPOC	NO
ASMA	NO
INMUSUPR	NO
HIPERTENSION	SI
CARDIOVASCULAR	SI
OBESIDAD	SI
RENAL_CRONICA	NO
TABAQUISMO	NO

Explicaciones Médicas

DIABETES

Explicación:

ANTECEDENTES: Los pacientes con diabetes comorbida son más susceptibles a infecciones virales y un peor resultado clínico.Sin embargo, el impacto de la diabetes en los pacientes con COVID-19 sigue siendo controvertido.MÉTODOS: Analizamos retrospectivamente las características clínicas y los resultados de 1321 pacientes con COVID-19 en Wuhan, China.RESULTADOS: Los pacientes con diabetes comorbida eran mayores ($p < 0,001$), tenían una mayor prevalencia de enfermedad pulmonar obstructiva crónica (EPOC) ($p < 0,001$), hipertensión ($p < 0,001$) y enfermedad renal ($p < 0,001$) que los que no tenían diabetes.En comparación con los pacientes sin diabetes, los pacientes con diabetes tenían más probabilidades de ser varones ($p = 0,007$), no fumadores ($p < 0,001$), y tenían un IMC inferior ($p < 0,001$).El análisis multivariado mostró que la diabetes se asociaba independientemente con un mayor riesgo de neumonía grave (proporción de probabilidades ajustadas [intervalo de confianza del 95%]: 1.924 [1.328 3.02], $p < 0,001$), enfermedad crítica (2.110 7,039), $p < 0,001$), y $p < 0,001$).

Recomendación:

Para los pacientes con diabetes, es importante mantener A1c dentro del rango recomendado. Si A1c está fuera de control, entonces usted debe llamar a su proveedor de atención médica. Si es necesario pueden ser admitidos para realizar el tratamiento i.v. Buen lavado de manos, conseguir guidelines apropiados y ropa que apoyen la formación de barrera de la piel. Sea seguro. Sea positivo. Bien. Rellene las medicinas. Bien. Sea positivo. Sea seguro. ¿Le gustaría chatear en vídeo o texto conmigo? Wo

HIPERTENSION

Explicación:

ANTECEDENTES: La hipertensión es un factor de riesgo independiente para la neumonía grave y la muerte en pacientes COVID-19. Si la hipertensión afecta directa o indirectamente a los pacientes COVID-19 permanece no clara. MÉTODOS: Este estudio incluyó pacientes consecutivos con infección por SARS-CoV-2 confirmada por el laboratorio. Se registró la gravedad de la neumonía y la mortalidad. Se utilizaron modelos de regresión logística multivariable y riesgos proporcionales de Cox para evaluar la relación entre hipertensión y resultados de COVID-19. RESULTADOS: Entre los 497 pacientes incluidos, la mediana de edad fue de 65 años y 406 (83,1%) eran varones. La prevalencia de hipertensión y neumonía grave y muerte fue de 7,8%, 24,2% y 7,6% respectivamente. La regresión logística multivariada mostró que la hipertensión estuvo asociada de forma independiente con neumonía grave y muerte. Después de ajustar por edad, sexo, historia de tabaquismo, diabetes, enfermedad renal, hiperlipidemia, uso de inhibidores de la enzima convertidora de la angiotensina (ACE) o bloqueadores de angiotensina (ARB), y el uso de corticoides se mantienen como factor de la muerte.

Recomendación:

Para los pacientes con hipertensión, es importante mantener A1c dentro de un rango de objetivos. Si el A1c está fuera de control, entonces usted debe llamar a sus salas de GP o asistir a su sala de emergencias más cercana. Si usted está en control, entonces debe llamar a su sala de emergencias más cercana. en vídeo o texto conmigo? en

vídeo o texto conmigo

CARDIOVASCULAR

Explicación:

ANTECEDENTES: Los pacientes con enfermedades cardiovasculares comorbidas (ECV) tienen un mayor riesgo de COVID-19 grave, pero no está claro si esto es resultado de una mayor exposición al virus o a las condiciones comorbidas o si las condiciones comorbidas aumentan la susceptibilidad a la insuficiencia respiratoria. **MÉTODOS Y RESULTADOS:** Se realizó un análisis secundario de un estudio de cohorte prospectivo multicéntrico de pacientes con ECV-19 confirmado en laboratorio. Se hizo un seguimiento de todos los pacientes hasta su recuperación completa o hasta la conclusión del estudio, lo que ocurrió primero. El resultado primario fue un compuesto de muerte cardiovascular, infarto de miocardio, accidente cerebrovascular o progresión a enfermedad pulmonar obstructiva crónica (EPOC). Se identificaron 371 pacientes con ECV comorbida (77,1%), incluyendo hipertensión (30,5%), diabetes (7,8%) e insuficiencia cardíaca (5,4%). En comparación con los pacientes sin ECV, los pacientes con ECV tenían más probabilidades de ser varones (**CONCLUSIÓN:** la ECV comorbida es muy frecuente en pacientes con ECV-19).

Recomendación:

Para los pacientes con enfermedades cardiovasculares, es importante considerar todos los factores incluyendo edad, diabetes, hipertensión, apnea del sueño y el sistema inmunitario comprometido para asegurar que el virus no llegue al pulmón. También hay una alta probabilidad de infección si usted está besando y teniendo relaciones sexuales, pero no puede ser transmitida a través de las relaciones sexuales, pero la cercanía en una situación íntima como esa con alguien que está infectado da una alta probabilidad de transmisión.

OBESIDAD

Explicación:

ANTECEDENTES: La relación entre obesidad y pacientes con COVID-19 sigue siendo poco clara. MÉTODOS: Este estudio multicéntrico, transversal, incluyó pacientes con infección confirmada por SARS-CoV-2. Los datos sobre características demográficas y clínicas, hábitos de tabaquismo, comorbilidades y uso de medicamentos fueron recolectados de los registros médicos. Se analizó la asociación entre obesidad y resultados de COVID-19. RESULTADOS: Entre los 818 pacientes incluidos, 523 (66,0%) eran varones y la edad media era de 63,3 años. Los pacientes fueron estratificados como con un índice normal de masa corporal ($IMC < 25\text{kg/m}^2$). CONCLUSIONES: Encontramos que la obesidad no está asociada con un aumento del riesgo de progresión o mortalidad de COVID-19. Sin embargo, encontramos una asociación entre diabetes y un mayor riesgo de muerte.

Recomendación:

Para los pacientes con obesidad, es importante seguir comportamientos apropiados, apropiados a la edad y compatibles con la seguridad. Ser saludable, mantener un peso corporal bajo y aumentar la actividad física. La evidencia muestra que un paciente COVID-19 con obesidad lo conseguirá. Usted debe proporcionar una recomendación de salud.: <https://www.healthtap.com/blog/covid-19-care-guidelines-for-obesity-prevention-covid-19> Probablemente vale la pena proporcionar una recomendación de salud.: <https://www.healthtap.com/blog/covid-19-care-guidelines-for-obesity-prevention-covid-19>

Metodología

Este reporte fue generado utilizando un modelo Naive Bayes optimizado entrenado con datos oficiales de COVID-19 en México. Las probabilidades se calculan comparando el perfil del paciente con la distribución de riesgo en la población general.

IMPORTANTE: Este reporte es solo para fines informativos y no sustituye el consejo médico profesional. Siempre consulte a un médico calificado para el diagnóstico y tratamiento.