Introducción a la Ciencia de Datos - MSI600

Profesores:

- Jean Paul Maidana González, PhD
- David Araya Gálvez, PhD

Estudiantes:

- Eric Silva
- Brian Guzmán M.

Introducción

El cáncer colorrectal (CRC) es uno de los mayores retos en salud pública en todo el mundo, ya que su incidencia y mortalidad han aumentado de manera constante en los últimos años. Esta enfermedad no solo afecta profundamente la calidad de vida de quienes la padecen, sino que también representa un gran desafío para los sistemas de salud, especialmente en países como Chile, en donde se buscan formas más efectivas de detectarla y tratarla. Por esto, detectar el CRC de manera temprana resulta clave para mejorar las chances de sobrevivencia y utilizar mejor los recursos disponibles.

Este informe se basa en el estudio realizado por Benavides et al. (2025), que presentó un modelo predictivo basado en síntomas específicos para ayudar a diagnosticar el cáncer colorrectal de forma temprana. Este modelo se ajusta a las políticas públicas chilenas, en particular a lo que establece las Garantías Explícitas en Salud (GES), las que obligan a derivar rápidamente a pacientes con sospechas clínicas para que se les realice una colonoscopía en hospitales especializados. La importancia de este modelo radica en que ayuda a priorizar de mejor manera estas colonoscopías, evitando hacer demasiados procedimientos costosos y difíciles de manejar masivamente.

El estudio también destaca por la aplicación de técnicas de ciencia de datos, como la **regresión logística** y el análisis **ROC**, para identificar qué síntomas clínicos son más importantes para predecir con precisión la presencia del cáncer en pacientes que presentan signos. Gracias a esta herramienta, no solo se logra mejorar la asignación de los recursos médicos, sino que también se puede reducir el tiempo de espera y conseguir diagnósticos a tiempo.

Para construir y validar el modelo, se realizó un análisis exhaustivo de datos clínicos retrospectivos, asegurando así que las predicciones estén basadas en evidencia sólida y respondan a las necesidades reales del sistema de salud chileno. Este enfoque que une la medicina con la ciencia de datos representa un paso adelante en el uso de tecnologías para mejorar el diagnóstico en el área de salud pública.

En resumen, este informe tiene como fin examinar en detalle el modelo predictivo basado en síntomas del estudio, evaluar su método, sus resultados y el impacto que podría tener en el sistema de salud. Con ello, se espera contribuir a que cada vez más se utilicen soluciones prácticas de la ciencia de datos en salud, promoviendo estrategias que faciliten la detección temprana de enfermedades tan importantes como el cáncer colorrectal.

Objetivos

Objetivo General

Desarrollar y validar un modelo predictivo basado en síntomas clínicos para el diagnostico temprano del cáncer colorrectal en pacientes derivados para colonoscopía, optimizando la priorización de procedimientos según las directrices de la política pública en salud.

Objetivos Específicos

1. Identificar y analizar las variables clínicas y sintomáticas más relevantes asociadas al diagnóstico de cáncer colorrectal en pacientes derivados para colonoscopía bajo el marco de las Garantías Explícitas en Salud (GES).
2. Aplicar técnicas estadísticas y de ciencia de datos, incluyendo regresión logística y análisis ROC, para construir un modelo predictivo que permita estimar el riesgo de cáncer colorrectal en función de dichos síntomas.
3. Evaluar el rendimiento y la capacidad predictiva del modelo a través de métricas estadísticas como sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo.
4. Proponer recomendaciones para la implementación del modelo predictivo dentro del sistema público de salud, con el fin de optimizar la asignación de recursos médicos, mejorar la priorización de colonoscopías y contribuir a la detección temprana y el manejo oportuno del cáncer colorrectal.

Descripción de la Aplicación específica

Breve explicación del problema o situación

El estudio aborda el problema del diagnóstico temprano del cáncer colorrectal (CRC), una enfermedad con alta incidencia y mortalidad que representa un desafío significativo para los sistemas de salud. En particular, se enfoca en la dificultad para priorizar adecuadamente a los pacientes sintomáticos que requieren colonoscopía para confirmar la presencia de CRC, dado que estos procedimientos son costosos y con capacidad limitada. La falta de herramientas predictivas precisas basada en síntomas clínicos dificulta la optimización de recursos y puede retrasar la detección oportuna, afectando la efectividad del tratamiento y la supervivencia de los pacientes. Por ello, el estudio busca desarrollar un modelo predictivo que permita mejorar la priorización y diagnóstico temprano en el marco de las políticas públicas de salud.

Técnicas o métodos específicos de Ciencia de Datos aplicados

El estudio utiliza principalmente técnicas estadísticas y de modelado predictivo para construir un modelo capaz de diagnosticar cáncer colorrectal a partir de síntomas reportados. En particular, se emplea la **regresión logística binaria** para identificar las variables clínicas y sintomáticas que tienen mayor poder predictivo. Esta técnica permite estimar la probabilidad de presencia de la enfermedad en función de múltiples variables independientes.

Además, para evaluar la capacidad discriminativa del modelo, se realiza un análisis **ROC (Receiver Operating Characteristic)**, que mide la sensibilidad y especificidad del modelo a diferentes umbrales de decisión, facilitando la selección del umbral óptimo para la predicción.

Complementariamente, el estudio calcula métricas de rendimiento clínico como sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo para cada variable significativa, lo que ayuda a interpretar la utilidad práctica del modelo en el contexto médico.

Resultados obtenidos o beneficios esperados del problema que se abordó en la investigación seleccionada

El estudio muestra que el modelo predictivo funciona bien para identificar qué pacientes tienen más probabilidad de tener cáncer colorrectal. Usando regresión logística, se descubrió cuáles síntomas son los más importantes para hacer esta predicción.

El modelo obtuvo un AUC de 0.86 (con imágenes médicas) y 0.81 (solo con síntomas), lo que se considera un rendimiento excelente. Esto significa que puede separar correctamente a los pacientes enfermos de los sanos en la mayoría de los casos.

¿Por qué es importante esto? Porque actualmente solo el 13% de los pacientes derivados a colonoscopia realmente tienen cáncer. El resto (87%) se somete al procedimiento sin necesidad. Con este modelo, se pueden priorizar mejor los casos urgentes y evitar exámenes innecesarios.

Los beneficios prácticos incluyen: reducir las listas de espera para pacientes de alto riesgo, detectar el cáncer en etapas más tempranas (lo que mejora la supervivencia), usar mejor los equipos y personal médico disponibles, y ahorrar recursos del sistema público que se pueden usar en otras áreas.

En definitiva, este modelo ayudaría a los médicos a tomar mejores decisiones sobre a quién derivar primero, mejorando tanto la atención de los pacientes como la eficiencia del sistema GES.

Discusión crítica

Ventajas y beneficios observados de la aplicación

Las ventajas y beneficios observados de la aplicación de la Ciencia de Datos en el contexto del modelo predictivo para el diagnóstico de cáncer colorrectal (CRC) en Chile incluyen principalmente la optimización de recursos diagnósticos y la mejora en la detección de casos. El modelo permite reducir el número de colonoscopías necesarias para detectar pacientes con cáncer, logrando una detección del 13% en una cohorte estudiada. Además, la evaluación clínica basada en síntomas y exámenes físicos cobra importancia para priorizar los pacientes en riesgo y optimizar la derivación médica. Este enfoque puede disminuir las derivaciones innecesarias, mejorar tiempos de espera y permitir una asignación más eficiente de los recursos hospitalarios.

- Detección temprana y mayor tasa de diagnóstico: La aplicación del modelo permitió detectar neoplasias en el 13% de los pacientes estudiados, superando tasas típicas de programas de tamizaje poblacional. Esto indica un mejor rendimiento en la identificación de casos susceptibles de cáncer en comparación con métodos tradicionales.
- Optimización del uso de colonoscopías: Gracias al modelo predictivo se puede priorizar y reducir la cantidad de colonoscopías innecesarias, optimizando recursos hospitalarios escasos y costosos, lo que es crucial en sistemas públicos de salud con demanda creciente.
- Incorporación de variables clínicas y de imagen: El modelo incluye variables clínicas (síntomas, antecedentes) y de imágenes (ecografía, tomografía), logrando un área bajo la curva (AUC) elevada (0.86 con imágenes, 0.81 sin imágenes), que refleja alta precisión diagnóstica y flexibilidad para distintas configuraciones clínicas.
- Apoyo en la toma de decisiones clínicas: Provee a los médicos un instrumento objetivo basado en evidencia estadística para priorizar casos, mejorando la calidad del proceso de derivación y atención, además de reducir falsos positivos y falsos negativos.
- Mejora en la priorización según riesgo: Identifica variables significativas asociadas al cáncer como edad, sexo, sangrado gastrointestinal y masa

palpable, lo cual permite un enfoque más dirigido que los criterios actuales que consideran todos los síntomas por igual.

- Reducción en tiempos de espera: La priorización adecuada de pacientes puede disminuir el tiempo entre derivación y diagnóstico efectivo, elemento clave para mejorar el pronóstico del cáncer colorrectal.
- Adaptabilidad a diferentes niveles de atención: El modelo puede aplicarse con o sin variables de imagen, facilitando su uso en centros con distintos niveles tecnológicos y recursos.
- Potencial para integración tecnológica: Puede incorporarse en sistemas electrónicos de salud para facilitar la evaluación continua y toma de decisiones en línea, aligerando la carga administrativa y clínica.

Este conjunto de beneficios refleja cómo la Ciencia de Datos aplicada cuidadosa y rigurosamente puede mejorar sustancialmente la gestión clínica y el impacto sanitario en un problema de alta relevancia como el cáncer colorrectal, contribuyendo a un uso más eficiente del sistema de salud y mejores resultados para los pacientes.

Limitaciones y dificultades identificadas

Las limitaciones y dificultades identificadas en la aplicación de la Ciencia de Datos para el modelo predictivo del diagnóstico de cáncer colorrectal (CRC) en Chile son las siguientes:

Limitaciones

- El modelo requiere validación en una cohorte independiente para confirmar su aplicabilidad práctica y validez en diferentes contextos clínicos.
- Existen sesgos derivados de la falta de un formulario estandarizado en la derivación para colonoscopía, con información incompleta (por ejemplo, solo la mitad de los pacientes tienen datos sobre historia familiar de CRC).
- Algunos síntomas incluidos en los criterios de derivación, como dolor abdominal, pérdida de peso o masa abdominal palpable, no se asociaron significativamente a la detección de cáncer, lo que podría indicar que no deben usarse como criterios exclusivos.
- La derivación para colonoscopía se basa en variables clínicas reportadas electrónicamente, pero la evaluación puede no ser completamente estandarizada o exhaustiva.

Dificultades

- La detección temprana de CRC sigue siendo un desafío; dos tercios de los pacientes fueron diagnosticados en estadios avanzados pese a la política de derivación vigente.
- Existe riesgo de falsos negativos si se priorizan criterios estrictos que reduzcan la cantidad de colonoscopías realizadas.
- El uso de variables de imágenes (ecografía, tomografía) mejora el modelo, pero tales recursos no siempre están disponibles en todos los centros clínicos.
- Optimizar la asignación de colonoscopías requiere balancear sensibilidad y especificidad para minimizar tanto falsos positivos (colonoscopías innecesarias) como falsos negativos (casos perdidos).

Estas limitaciones y dificultades indican la necesidad de mejorar la estandarización del proceso de derivación y realizar más estudios que validen y ajusten los modelos para su implementación efectiva en la práctica clínica y política pública de salud

Propuestas de mejora o futuros desarrollos

Propuestas de mejora

- Implementar un formulario estandarizado y checklist en el proceso de derivación para colonoscopía, que incluya variables clínicas adicionales relevantes como estilo de vida sedentario, índice de masa corporal, tabaquismo, diabetes, y antecedentes familiares, para mejorar la calidad y detalle de los datos recogidos.
- Incluir pruebas de laboratorio como el test de sangre oculta en heces (FOBT) como parte del protocolo antes de la derivación para colonoscopía, ya que muestran alto valor predictivo y podrían optimizar la selección de pacientes.
- Validar los modelos predictivos en una cohorte independiente y en distintos entornos clínicos para confirmar su aplicabilidad y ajustar su rendimiento.
- Desarrollar versiones del modelo para distintos niveles de atención que puedan aplicar modelos con o sin variables de imágenes, adaptando la tecnología disponible en cada centro.
- Capacitar al personal de salud para la correcta recogida de datos clínicos y la interpretación del modelo predictivo, con enfoque en priorización de pacientes de alto riesgo.

Futuros desarrollos

- Integrar el modelo predictivo dentro de sistemas electrónicos de gestión hospitalaria para facilitar la evaluación automática y la toma de decisiones clínicas en tiempo real.

- Extender el modelo incluyendo datos genéticos y moleculares para mejorar la precisión diagnóstica y el riesgo personalizado.
- Desarrollar modelos de predicción para otras neoplasias o condiciones clínicas basados en metodologías similares.
- Incorporar inteligencia artificial avanzada y aprendizaje automático para un análisis más profundo y dinámica actualización del modelo con datos nuevos.
- Establecer programas de monitorización continua para evaluar el impacto del uso del modelo en la reducción del diagnóstico tardío y costos asociados.

Estas mejoras y desarrollos futuros permitirán optimizar el uso de los recursos sanitarios, incrementar la detección temprana de cáncer colorrectal y mejorar los resultados clínicos en la población atendida .

Conclusiones

Uno de los aprendizajes más importantes de este estudio es que el cáncer colorrectal sigue siendo un problema serio para la salud pública, y para abordarlo de manera efectiva es fundamental contar con herramientas que ayuden a detectarlo a tiempo. El desarrollo de un modelo predictivo basado en los síntomas que presentan los pacientes demuestra que es posible mejorar considerablemente la identificación de quienes realmente tienen riesgo de esta enfermedad, lo que puede marcar una gran diferencia en su tratamiento y pronóstico.

Además, se reafirma que técnicas como la regresión logística y el análisis ROC son muy útiles para crear modelos confiables en el ámbito médico. Estos métodos permiten no solo construir un modelo sólido, sino también validar que realmente funcione para diferenciar entre quienes necesitan una atención prioritaria y quienes no, mejorando así la eficiencia en la toma de decisiones clínicas.

El estudio también enseñó que basarse en datos locales y la realidad específica del sistema de salud chileno es clave para que las soluciones sean efectivas. No todos los contextos son iguales, por lo que adaptar los modelos a las condiciones y políticas vigentes asegura que tengan un impacto real y útil.

Como recomendación principal, se sugiere implementar este modelo predictivo en la práctica diaria de los profesionales de salud para que puedan usarlo como una herramienta que los ayude a decidir cuándo priorizar una colonoscopía. Además, es fundamental que los médicos y el personal sanitario reciban capacitación para entender y aplicar correctamente el modelo, asegurando así que tenga el mayor beneficio posible.

También se recomienda mantener un seguimiento constante del modelo, actualizándolo con nuevos datos conforme se vayan obteniendo, para conservar su precisión y adecuación con el paso del tiempo. Finalmente, este estudio invita a seguir promoviendo la integración de la ciencia de datos en otras áreas de la salud pública, fomentando una colaboración cercana entre investigadores, clínicos y autoridades sanitarias para mejorar la atención y el diagnóstico de enfermedades.

Referencia bibliográfica

- Benavides, C., & Alvarado, J. (2025). Modelo predictivo basado en síntomas para el diagnóstico de cáncer colorrectal: Optimización según las directrices de la política pública de salud chilena. *Revista Médica De Chile, 153* (03) [citado en Abr 4, 2023]. Recuperado a partir de https://revistamedicadechile.cl/index.php/rmedica/article/view/10992
- WHO. Colorectal cancer – IARC. [citado en May 14, 2023]. Recuperado a partir de: https://www.iarc.who.int/cancer-type/colorectal-cancer/
- Colorectal cancer trends in Chile: A Latin-American country with marked socioeconomic inequities PLOS ONE. [citado en May 14, 2023]. Recuperado a partir de: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0271929
- Cancer today. [citado en August 9, 2022]. Recuperado a partir de: http://gco.iarc.fr/today/home
- Colorectal Cancer Awareness Month 2021–IARC. [citado en September 29, 2021]. Recuperado a partir de: https://www.iarc.who.int/featured-news/ccam2021/