“PREDICCIÓN DEL VALOR MONETARIO DE LAS ATENCIONES DE PERSONAS CON DIAGNÓSTICO DEFINITIVO DE DIABETES MELLITUS”

Autor/Autores: **Robbie Shun Ching Mamani Uruchi, Jhossep Pedro Gómez Mercado, Jorge Luis Campana Vivanco, Víctor André Pozo Cisneros**

Resumen- En el presente informe se describe una metodología detallada para predecir la cantidad de dinero que el Seguro Integral de Salud deberá desembolsar en atenciones médicas para los afiliados con un diagnóstico definitivo de Diabetes Mellitus. La metodología propuesta incluye la adquisición del dataset, el preprocesamiento de datos, la implementación y evaluación del modelo predictivo.

1. Introducción

Descripción del problema:
La diabetes mellitus se configura como una enfermedad crónica que afecta a una cantidad considerable de individuos en todo el mundo, constituyendo una carga significativa tanto para los pacientes como para los sistemas de salud. Según la Federación Internacional de Diabetes (IDF), en 2021 se estima que alrededor de 537 millones de adultos viven con diabetes, y se proyecta un aumento a 643 millones para 2030 y 783 millones para 2045 (International Diabetes Federation, 2021).

En el contexto peruano, la situación no difiere en términos de preocupación. Según datos proporcionados por el Ministerio de Salud del Perú, aproximadamente 2.3 millones de peruanos sufren de diabetes. La asignación presupuestaria para la atención de personas con diabetes en el sistema de salud pública del país ha sido objeto de debate recurrente. En 2022, el presupuesto anual destinado a la atención de la diabetes ascendió a aproximadamente 19 millones de soles, equivalentes a unos 5 millones de dólares estadounidenses (Ministerio de Salud del Perú, 2022).

Walter Díaz, en un artículo publicado en Infobae, expresa que el presupuesto asignado al sector salud es insuficiente, lo cual conduce a una atención deficiente de los pacientes con diabetes en Perú. La falta de comprensión sobre los costos asociados con el tratamiento de la diabetes resulta en una gestión inadecuada de los recursos destinados a este sector (Díaz, año).

Hipótesis y/o pregunta a abordar:
Con información proporcionada de la Plataforma Nacional de Datos Abiertos, es posible predecir con precisión el valor neto de las atenciones médicas que el Seguro Integral de Salud (SIS) deberá desembolsar para pacientes con Diabetes Mellitus mediante el uso de modelos de inteligencia artificial, utilizando datos demográficos y clínicos disponibles.

Objetivos:
- Desarrollar un modelo de inteligencia artificial capaz de predecir el valor neto que gasta el SIS en cada paciente.
- Identificar y seleccionar las características más relevantes que influyen en los costos de atención médica de los pacientes con Diabetes Mellitus.
- Evaluar y optimizar el rendimiento del modelo predictivo para asegurar su precisión y aplicabilidad en un entorno real.
- Proveer recomendaciones para la asignación eficiente de recursos en el sistema de salud pública, basadas en las predicciones del modelo.

2. Trabajos relacionados

Predicción del riesgo de diabetes tipo 2 y su evaluación de efectos utilizando el modelo XGBoost:
Investigadores chinos desarrollaron un modelo de predicción del riesgo de diabetes tipo 2 (DT2) basado en el algoritmo XGBoost, utilizando datos de 100,000 pacientes. El modelo XGBoost superó a otros modelos como la regresión logística y las máquinas de vectores de soporte en la precisión de la predicción, alcanzando un área bajo la curva (AUC) de 0.954. Además, demostró ser eficaz para identificar individuos de alto riesgo con un valor predictivo positivo (PPV) del 85.2% y una sensibilidad del 90.3%. Esta herramienta prometedora puede ayudar a identificar a las personas con mayor riesgo de desarrollar DT2 para intervenciones preventivas tempranas.

https://www.mdpi.com/2227-9032/8/3/247

Modelo de Predicción de Costos en Servicios de Salud Soportado en Simulación Discreta
En este trabajo, se presenta un modelo de simulación discreta para estimar el costo de prestación de servicios de salud en una Entidad

Prestadora de Servicios de Salud (EPS). El modelo considera variables relevantes como la distribución de servicios, el número de servicios por paciente y el costo por servicio. Se simularon cuatro escenarios para evaluar el impacto de estas variables en el costo total.

Los resultados muestran que la predicción del costo total utilizando promedios puede tener variaciones significativas en comparación con la simulación. Este modelo permite a las EPS estimar el costo de prestación de servicios de salud con mayor precisión, brindando información valiosa para la toma de decisiones.

Además, puede ser utilizado para analizar diferentes escenarios de operación del sistema, permitiendo evaluar el impacto de cambios en las variables de decisión sobre el costo total. Es importante contar con datos históricos para alimentar el modelo de simulación. El análisis de varianza (ANOVA) puede ser utilizado para identificar las variables de decisión que más impactan en el costo.
https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642014000400019

Costos de enfermedades: clasificación y perspectivas de análisis
El análisis exhaustivo de los costos asociados a las enfermedades es fundamental para la toma de decisiones eficientes en el ámbito de la salud. Esta investigación revisa la perspectiva de análisis en los estudios de costos de enfermedades, centrándose en cómo esta perspectiva puede influir en la estimación de los costos de una condición de salud.

Los estudios revisados sugieren que la perspectiva de análisis puede variar significativamente, con implicaciones directas en los resultados obtenidos. Las perspectivas más comunes son la social, la del proveedor de servicios de salud, la del asegurador o tercer pagador, y la del paciente. Cada una de estas perspectivas considera diferentes aspectos de los costos asociados a una enfermedad.

Desde la perspectiva social, se busca evaluar el impacto global de una condición de salud, teniendo en cuenta no solo los costos directos y no sanitarios, sino también los relacionados con la muerte, la discapacidad, el sufrimiento y el dolor. En contraste, la perspectiva del proveedor de servicios de salud se centra principalmente en los costos asociados a la atención sanitaria, incluyendo también los recursos destinados al gerenciamiento de los servicios.

Por otro lado, desde la perspectiva del tercer pagador, los costos relevantes están vinculados al proceso de salud-enfermedad-atención, ya que estos determinan la cotización de las pólizas de riesgo y el margen de ganancia esperada. Finalmente, desde la perspectiva del paciente, se destacan los costos no sanitarios, los costos indirectos y los intangibles, especialmente si se pretende estimar la repercusión económica de una pérdida en el estado de salud.
https://www.redalyc.org/journal/562/56249528005/html/

3. Metodología

- Enfoque(s) propuesto: Para predecir el valor neto que el Seguro Integral de Salud (SIS) deberá desembolsar en atenciones médicas para los afiliados con un diagnóstico definitivo de Diabetes Mellitus, se propone un enfoque basado en el uso de técnicas de Machine Learning (ML). El enfoque incluye las siguientes etapas:

1. Adquisición y preprocesamiento de datos.
2. Análisis exploratorio de datos (EDA).
3. Selección de características.
4. División de datos en conjuntos de entrenamiento y prueba.
5. Desarrollo y entrenamiento del modelo predictivo.
6. Evaluación y optimización del modelo.
7. Implementación y monitoreo del modelo.

Descripción formal del problema:
El problema puede ser descrito formalmente de la siguiente manera:

Conjunto de Datos: $D = \{(x1,y1),(x2,y2),...,(xn,yn)\}$
donde xi representa el vector de características del i-ésimo afiliado y yi es el valor neto de las atenciones médicas que se desea predecir.

Vector de Características:
$xi = [xi1,xi2,...,xim]$, donde xij es la j-ésima característica del i-ésimo afiliado.

Variable Objetivo: yi es el costo en soles (moneda peruana) que el SIS debe desembolsar por el i-ésimo afiliado.

Función Predictiva:
$f: Rm \rightarrow R$, donde $f(xi) \approx yi$
.
Comportamiento entrada/salida del enfoque:
Entrada: Vector de características xi de un afiliado, que incluye datos demográficos, diagnósticos adicionales, y detalles de atenciones médicas previas.

Salida: Valor neto predicho $yi^$ que el SIS deberá desembolsar por las atenciones médicas del afiliado.

Viabilidad del proyecto:

Se tiene el dataset de la Plataforma Nacional de Datos Abiertos del Perú, que entre sus principales características se tienen:

EDAD	Edad del afiliado a la fecha de corte de afiliados activos al SIS
UBIGEO	Ubigeo del establecimiento de adscripción del afiliado
SEXO	Sexo del afiliado: Femenino, Masculino
TIPO_DIABETES	Tipo de Diabetes Mellitus según diagnóstico registrado
CON_DX_OBESIDAD	Indica si registra al menos un diagnóstico definitivo de Obesidad/Dislipidemia desde el 2018 a la fecha de evaluación
CON_DX_HIPERTENSION	Indica si registra al menos un diagnóstico definitivo de Hipertensión desde el 2018 a la fecha de evaluación
CON_DX_SALUDMENTAL	Indica si registra al menos un diagnóstico definitivo de salud mental desde el 2018 a la fecha de evaluación
CANT_ATENCIONES	Cantidad de atenciones recibidas por el afiliado en el último trimestre con al menos un diagnóstico de Diabetes Mellitus (DM).
VALOR_NETO	Valor reconocido por el SIS del consumo de la atención. Expresada en moneda peruana llamada sol
CANT_ATENCIONES_HOSP	Cantidad de atenciones por hospitalización recibidas por el afiliado en el último trimestre.
VALOR_NETO_HOSP	Valor neto de las atenciones recibidas por el afiliado por hospitalización en el último trimestre
DIAS_HOSP	Día(s) de hospitalización del afiliado en el último trimestre
NIVEL_ULT_ATE	Nivel del establecimiento de la última atención recibida por el afiliado

4. Experimentación y Resultados

- Setup experimental:

1. Describir datos usados (o método para obtenerlos) (si aplica).

A partir del dataset inicial, creamos un dataframe en donde, tras un preprocesado de información, se tienen 69629 filas y 17 columnas. Entre las columnas se tienen: EDAD, FECHA_PRIMER_DX, CANT_ATENCIONES, VALOR_NETO, CANT_ATENCIONES_HOSP, VALOR_NETO_HOSP, DIAS_HOSP, TIPO_DIABETES_Diabetes mellitus no especificada, TIPO_DIABETES_Diabetes mellitus asociada con desnutrición, TIPO_DIABETES_Diabetes mellitus tipo 1, TIPO_DIABETES_Diabetes mellitus tipo 2, TIPO_DIABETES_Otras Diabetes mellitus especificada, SEXO_FEMENINO, SEXO_MASCULINO, CON_DX_OBESIDAD_SI, CON_DX_HIPERTENSION_SI, CON_DX_SALUDMENTAL_SI

2. Los datos serán evaluados con las siguientes métricas: Mean squared error, Mean absolute error, Explained variance score, R2 score. A continuación, cada uno será explicado:

   El Mean Squared Error (MSE) es una medida de la calidad de un modelo de regresión. Se calcula tomando la media de los cuadrados de los errores o diferencias entre los valores predichos por el modelo y los valores reales.

   El Mean Absolute Error (MAE) mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Es la media de los valores absolutos de las diferencias entre los valores predichos y los valores reales.

   El Explained Variance Score mide la proporción de la varianza total de los datos que es explicada por el modelo. Su valor varía entre 0 y 1, donde 1 indica que el modelo explica toda la varianza en los datos y 0 indica que el modelo no explica ninguna varianza.

   El R2 Score o Coeficiente de Determinación es otra medida de la calidad del ajuste de un modelo de regresión. Indica la proporción de la varianza de la variable dependiente que es explicada por las variables independientes en el modelo.

3. Los componentes principales incluyeron modelos de regresión lineal (Linear Regression, Ridge Regression, Lasso Regression, ElasticNet) para explorar diferentes enfoques de ajuste y regularización.}

- Resultados y Discusión:

A continuación se mostrarán los resultados experimentales para seleccionar el algoritmo de regresión más adecuado.

BoxPlots de todos los algoritmos

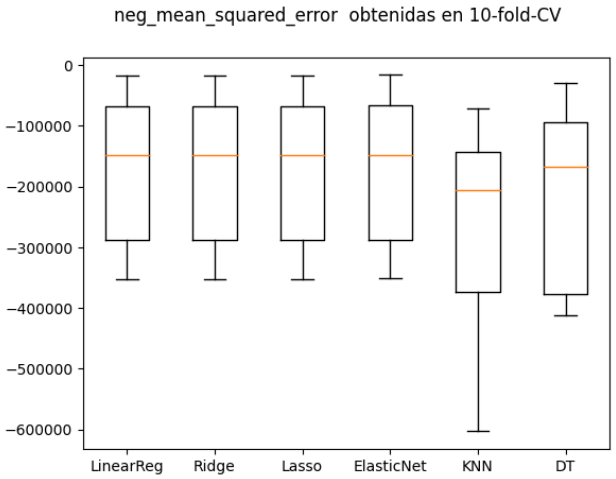


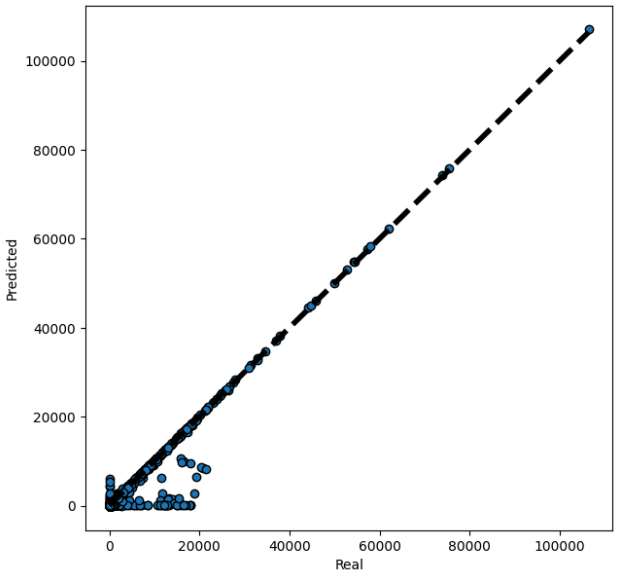
Tabla de los errores cuadráticos medios

Algoritmo	Métrica
LinearReg	-171049.52277501527 (122323.53637544796)
Ridge	-171049.06730565123 (122323.6583053219)
Lasso	-170982.17199240482 (122306.38817449252)
ElasticNET	-170725.18517104874 (122738.627022963)
KNN	-170725.18517104874 (122738.627022963)
DTT	-216948.12802309898 (144772.31123505672)

De acuerdo a la tabla, se puede observar que el mejor algoritmo es ElasticNet el cual tiene un promedio de -170725.18517104874. Además, en el boxplot se aprecia que es el más robusto y cercano a 0. Evaluando el modelo con el algoritmo ElasticNet para la data separada para "test" se obtienen los siguientes parámetros:

Mean squared error	166903.2196948042
Mean absolute error	87.96188475837053
Explained variance score	0.9276173739786061
R2 score	0.9276145773705329

Según el estadístico Jim Frost, un R2 score es considerablemente alto si este es mayor al 90%. El R2 score evidenciando en la tabla presenta un valor de 0.9276145773705329 lo cuál nos indicaría que se trata de un buen modelo.



Como los valores numéricos del valor neto son muy grandes, habrá valores muy alejados de la recta y que el modelo difícilmente podrá predecir. Por ejemplo, para valores por debajo de 20000, el modelo no es tan adecuado para predecir. Los puntos tienden a estar más cercanos cuando son valores mayores a 20000, ahí es donde el modelo predice prácticamente bien como se observa en el resultado del R2 score.

5. Conclusión

Se ha determinado que ElasticNet es el algoritmo de regresión más apropiado para predecir valores netos, basándose en los datos experimentales obtenidos. Este hallazgo subraya la capacidad de ElasticNet para modelar con precisión variables complejas y proporcionar predicciones robustas en escenarios de análisis predictivo.

En consecuencia, mediante este algoritmo es posible estimar de manera aproximada los futuros gastos del seguro de personas con diabetes. Este resultado ofrece una herramienta útil para la planificación financiera y la gestión de riesgos en el contexto de la salud pública y los seguros médicos.

Al desarrollar un modelo de inteligencia artificial capaz de predecir el 92.76% de los datos se concluye que el algoritmo ElasticNet logró un modelo fuerte debido a su valor de R2 Score.

Al realizar la matriz de correlación se muestra que el valor neto tiene alta correlación con los días en el hospital, por lo cual se concluye que esta es la característica que más influye en los costos de atención médica.

6. Sugerencias de trabajos futuros

- Explorar técnicas avanzadas de Machine Learning o Deep Learning para mejorar la precisión del modelo en la predicción de costos de atención médica. Esto podría incluir el uso de redes neuronales profundas o modelos de atención que puedan capturar relaciones más complejas entre las variables.

- Investigar la inclusión de variables adicionales, como hábitos de vida, características genéticas específicas, o datos socioeconómicos, para ver cómo afectan los costos de atención médica y mejorar así la precisión del modelo.

- Realizar validaciones externas del modelo utilizando datos de otras fuentes o de diferentes períodos temporales para verificar su robustez y generalización en diferentes contextos y condiciones.

7. Implicancias éticas

- El manejo de datos de salud altamente sensibles conlleva riesgos significativos de violaciones de privacidad si no se gestiona adecuadamente. La exposición de información médica podría tener graves consecuencias para los pacientes. Es esencial implementar medidas de seguridad robustas, como el cifrado de datos tanto en reposo como en tránsito, controlar estrictamente el acceso a los datos y asegurar que solo el personal autorizado pueda manipularlos. Además, es importante cumplir con regulaciones de protección de datos como la GDPR, y anonimizar los datos siempre que sea posible para proteger la identidad de los pacientes.

- Los sistemas que manejan grandes volúmenes de datos de salud son objetivos atractivos para ciberataques, lo que podría resultar en la exposición de información sensible. Para prevenir esto, se deben realizar auditorías de seguridad periódicas y pruebas de penetración para identificar y corregir vulnerabilidades. Es crucial implementar controles avanzados de acceso, como la autenticación multifactorial, y capacitar al personal sobre buenas prácticas de ciberseguridad para prevenir ataques de ingeniería social, como el phishing, asegurando así la protección continua de los datos.

- Las decisiones derivadas de un modelo predictivo deben ser comprensibles y explicables para mantener la confianza tanto de los pacientes como de los profesionales de la salud. Utilizar modelos interpretables o desarrollar técnicas que permitan explicar las predicciones de modelos más complejos es fundamental. Proveer explicaciones claras sobre cómo se generan las predicciones y los factores considerados ayudará a garantizar la transparencia. Involucrar a expertos en salud en el desarrollo y validación del modelo asegura que las predicciones sean clínicamente relevantes y precisas, promoviendo así la confianza en su uso.

- Las predicciones del modelo podrían influir significativamente en las decisiones de tratamiento y en la asignación de recursos, afectando potencialmente la calidad de la atención que reciben los pacientes. Es crucial que las predicciones se utilicen como herramientas complementarias y no como la única base para la toma de decisiones. Proveer formación a los profesionales de la salud sobre cómo interpretar y utilizar las predicciones de manera ética y efectiva es esencial. Además, implementar un sistema de revisión y retroalimentación para ajustar el modelo basado en los resultados clínicos y la satisfacción del paciente puede ayudar a asegurar un impacto positivo y equitativo en la atención médica.

8. Link del repositorio del trabajo

https://github.com/JhossepGM/Proyecto_IA

9. Declaración de contribución de cada integrante

Integrante	Aporte
Jorge Luis Campana Vivanco	Elaboración del código y redacción del informe
Jhossep Pedro Gómez Mercado	Elaboración del código y redacción del informe
Robbie Shun Ching Mamani Uruchi	Elaboración del código y redacción del informe
Víctor André Néel Pozo Cisneros	Elaboración del código y redacción del informe

10. Referencias

- International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th edition. Retrieved from https://diabetesatlas.org/
- Ministerio de Salud del Perú. (2021). Estadísticas de Diabetes en el Perú. Retrieved from https://www.gob.pe/minsa
- Diario Gestión. (2023). Presupuesto asignado para la diabetes en Perú. Retrieved from https://gestion.pe
- Defensoría del Pueblo del Perú. (2022). Informe de brechas en la atención de enfermedades crónicas. Retrieved from Defensoría del Pueblo
- "How High Does R-squared Need to Be?" Statistics By Jim. https://statisticsbyjim.com/regression/how-high-r-squared/
-