THE BRIDGE DIGITAL TALENT ACCELERATOR

MACHINE LEARNING PROJECT

MODELO DE PREDICCIÓN DE PRECIOS SEGUROS MÉDICOS

DATA SCIENCE ONLINE 2025

Presentado por:
Deivid Jimenez Sanchez

ÍNDICE

CONTENIDO DEL INFORME

RESUMEN EJECUTIVO

Este proyecto tiene como objetivo desarrollar un modelo de machine learning capaz de predecir el costo del seguro médico de una persona en función de variables como la edad, el índice de masa corporal (IMC), el número de hijos, el género, la región de residencia y si es fumador o no. Para ello, se ha utilizado un conjunto de datos públicos y se aplicaron diversas técnicas de regresión, incluyendo Regresión Lineal, Random Forest y Gradient Boosting.

El proceso incluyó la exploración y visualización de los datos, el preprocesamiento mediante pipelines, la codificación de variables categóricas y la normalización de los datos numéricos. Posteriormente, se entrenaron y evaluaron los modelos utilizando métricas como MAE, MSE y R^2 Score, lo que permitió comparar su rendimiento y seleccionar el más adecuado. Finalmente, el modelo fue guardado en formato .joblib para su reutilización

El resultado es un sistema de predicción eficiente y reutilizable que permite estimar con buena precisión los costos médicos individuales, lo cual puede ser útil para aseguradoras, instituciones médicas o propósitos educativos relacionados con análisis predictivo y ciencia de datos.

Nota: Los datos de este proyecto son totalmente ficticios, extraídos de Kaggle por lo cual no hacen referencia a ninguna empresa aeguradora real

INTRODUCCIÓN

El análisis predictivo aplicado al sector salud se ha convertido en una herramienta clave para la toma de decisiones estratégicas. En este proyecto, se aborda el problema de predecir el costo del seguro médico a partir de variables demográficas y de estilo de vida. Utilizando técnicas de machine learning, se busca construir modelos que permitan estimar dichos costos con precisión, identificando patrones en los datos que influyen significativamente en los precios de las primas médicas.

Este proyecto aplica técnicas de machine learning para predecir el costo del seguro médico en función de variables personales como edad, IMC, número de hijos y hábitos de salud. El objetivo es construir modelos predictivos precisos que ayuden a entender cómo estas variables influyen en el costo, con posibles aplicaciones en el sector asegurador y la analítica de salud.

OBJETIVOS

Definición objetivo general

Desarrollar un modelo de machine learning capaz de predecir el costo del seguro médico a partir de variables demográficas y de salud, utilizando técnicas de análisis de datos y regresión.

Definición objetivos específicos

- Analizar y explorar el conjunto de datos de seguros médicos para comprender las variables disponibles.
- Preprocesar los datos mediante codificación, normalización y división en conjuntos de entrenamiento y prueba.
- Entrenar y comparar diferentes modelos de regresión (Lineal, Random Forest y Gradient Boosting).
- Evaluar el rendimiento de los modelos utilizando métricas como MAE, MSE y R^2.
- Guardar el modelo más eficiente para su uso posterior en aplicaciones predictivas.

DESCRIPCIÓN DE LOS ANÁLISIS REALIZADOS

Para alcanzar los objetivos del proyecto, se realizaron diversas etapas de análisis de datos:

El análisis comenzó con una exploración del conjunto de datos, identificando la distribución de las variables y la existencia de posibles correlaciones. Se aplicaron técnicas de visualización con gráficos de dispersión, histogramas y mapas de calor para comprender las relaciones entre las variables independientes y el costo del seguro médico.

Posteriormente, se llevó a cabo un preprocesamiento que incluyó la codificación de variables categóricas mediante OneHotEncoding, la estandarización de variables numéricas y la división del conjunto de datos en entrenamiento y prueba. Se implementaron pipelines para automatizar este flujo de preparación.

A continuación, se entrenaron y compararon distintos modelos de regresión: Regresión Lineal, Random Forest y Gradient Boosting. Cada modelo fue evaluado utilizando métricas de desempeño como el error absoluto medio (MAE), el error cuadrático medio (MSE) y el coeficiente de determinación (R^2), permitiendo seleccionar el modelo más preciso. Finalmente, el modelo elegido fue almacenado en un archivo .joblib para facilitar su uso posterior sin necesidad de reentrenamiento.

RESULTADOS

Los modelos entrenados mostraron distintos niveles de precisión al predecir el costo del seguro médico. La Regresión Lineal presentó un rendimiento aceptable como punto de partida, pero modelos más complejos como el Random Forest y Gradient Boosting ofrecieron mejores resultados en términos de precisión.

El modelo de Gradient Boosting fue el que obtuvo el mejor desempeño, con el menor error absoluto medio (MAE) y el mayor coeficiente de determinación (R^2), lo que indica una mayor capacidad para capturar la variabilidad en los datos. Estos resultados sugieren que los métodos de ensamblado son más adecuados para este tipo de problema no lineal.
En general, el análisis demostró que variables como el IMC, el hábito de fumar y la edad tienen un impacto significativo en el costo del seguro, confirmando la utilidad del modelo tanto para análisis exploratorio como para predicción práctica en contextos reales.

Modelo	MAE	RMSE	R^2
Regresión Lineal	3888.44	7814.06	0.6067
Random Forest	1951.56	3693.63	0.8482
Gradient Boosting	2029.37	4384.07	**0.8762**

CONCLUSIONES

- El análisis predictivo mediante machine learning permite estimar con buena precisión el costo del seguro médico, aportando una herramienta útil tanto para aseguradoras como para estudios de riesgo y salud.

- Las variables más influyentes en el costo fueron el IMC, el hábito de fumar y la edad, lo que valida la lógica del modelo y su alineación con factores reales de riesgo médico.

- Aunque la Regresión Lineal ofreció una base comprensible y rápida de implementar, su capacidad predictiva fue limitada frente a métodos más avanzados.

- En contraste, los modelos Random Forest y Gradient Boosting demostraron mejor desempeño, siendo este último el más preciso.

- El uso de pipelines de preprocesamiento y la modularización del flujo de trabajo mejoraron la reproducibilidad del proyecto, haciendo que el modelo final pueda reutilizarse fácilmente en entornos reales o integrarse en una aplicación.

- Finalmente, este proyecto demuestra que, con un conjunto de datos bien estructurado y técnicas adecuadas, es posible construir soluciones inteligentes que optimicen procesos de toma de decisiones en sectores como la salud y los seguros.

RECOMENDACIONES O POSIBLES MEJORAS A FUTURO

- Incorporar más variables predictoras, como historial médico, actividad física o antecedentes familiares, podría mejorar la precisión del modelo y hacerlo más representativo de escenarios reales.

- Ampliar el conjunto de datos con información más actual o de diferentes regiones geográficas permitiría entrenar modelos más robustos y generalizables.

- Fomentar el uso de herramientas de análisis de datos en el sector salud, ya que permiten tomar decisiones más informadas, identificar riesgos y personalizar las políticas de seguros de forma justa y eficiente.

- Implementar una interfaz de usuario o API permitiría que el modelo sea accesible de forma interactiva por otros usuarios o sistemas externos.

- Monitorear el rendimiento del modelo en producción para detectar posibles degradaciones a lo largo del tiempo y actualizarlo según sea necesario.

- Presentar los resultados de forma clara y visual a personas no técnicas para facilitar su comprensión y promover la adopción de soluciones basadas en datos