

Ciencia de datos

Módulo 6: Aprendizaje de máquina supervisado

Tarea 1

Objetivo: *Hacer un modelo de regresión que prediga cuánto se le debe cobrar a un paciente por una aseguradora dado su historial clínico básico e información demográfica.*

Instrucciones:

1. Descargar el dataset “insurance.csv” de la página: <https://www.kaggle.com/mirichoi0218/insurance>
2. Responder las preguntas listadas abajo en un notebook de Python. Cada pregunta tiene una puntuación, en total suman 10.
3. Exportar el notebook en formato PDF y subirlo a Moodle. Solo se aceptarán entregas en PDF.

Preguntas:

1. ¿Existen valores nulos en alguna columna? (1 punto)
2. Graficar la distribución de la columna “charges” (1 punto)
3. Graficar la distribución conjunta de “charges” y “age” para los no fumadores (1 punto)
4. Graficar la matriz de correlación entre las variables incluyendo la variable “charges” (1 punto)
5. Ajustar los siguientes modelos de regresión y encontrar las métricas de R^2 para cada uno de ellos (se debe predecir la variable “charges”) (2 puntos):
 - a. Linear Regression
 - b. Ridge Regression
 - c. Lasso Regression
 - d. KNN regression
6. ¿Cuáles son las mejores variables para la predicción de “charges”? (1 punto)
7. Discuta cuál modelo entrega mejor predictibilidad y por qué (2 punto)
8. Haga un resumen ejecutivo de los resultados del modelo para presentar a una audiencia de negocio (1 punto)

Elaboro: M.Sc. Favio Vázquez

Calificación: 35% del módulo