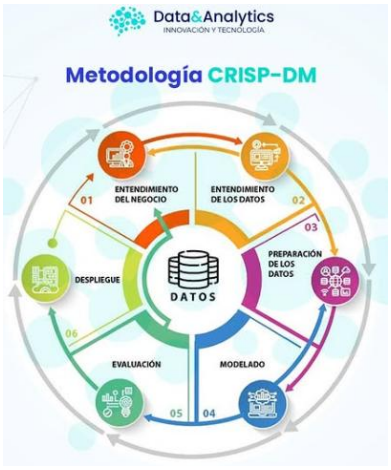
Proceso de Machine Learning aplicado a datos de salud - INEGI



Los datos provienen de la encuesta ENSANUT 2018 del INEGI, del cuestionario de etiquetado frontal de alimentos.
La intensión es predecir si un mexicano va a leer y ocupar la información que muestra el etiquetado frontal en alimentos empacados y bebidas embotelladas.
En el diccionario de datos se encuentra la descripción de las columnas y contempla:

- Datos sociodemográficos
- Preguntas del cuestionario

1. En promedio, ¿cuántas calorías considera usted que debe consumir una persona sana, de su misma edad y sexo, en un día?

CRUZA UN CÓDIGO

Menos de 500 calorías	1	De 2 001 a 3 000 calorías	5
De 500 a 1 000 calorías	2	De 3 001 a 4 000 calorías	6
De 1 001 a 1 500 calorías	3	Más de 4 000 calorías	7
De 1 501 a 2 000 calorías	4	No sabe / No responde	9

3. ¿Usted sabe si los alimentos empacados y las bebidas embotelladas tienen información sobre su contenido nutrimental?

CRUZA UN CÓDIGO

SÍ	1	
NO	2	Pasa a 7
No sabe / No responde	9	

4. ¿Usted lee la información nutrimental de los alimentos empacados y las bebidas embotelladas que compra?

CRUZA UN CÓDIGO

SÍ	1	
NO	2	Pasa a 7
No sabe / No responde	9	

7. Cuando compra alimentos empacados y/o bebidas embotelladas, ¿con qué frecuencia elige un producto por la información de los empaques? (logotipos o leyendas de salud)

CRUZA UN CÓDIGO

Nunca	1
Casi nunca	2
A veces	3
Casi siempre	4
Siempre	5
No sabe / No responde	9

1. Entendimiento del negocio
   a. En un párrafo, haz una descripción de la razón e importancia por la cual el Gobierno de México creó el etiquetado frontal de alimentos. Incluir cuándo entró en vigor.
   b. Crear una tabla con las variables, su tipo, sus posibles valores, unidades (si aplica) y una definición sencilla.
   c. Plantear los problemas de clasificación iniciales (omitir distribución de probabilidad), considerando lo siguiente:
      i. Problema 1: Predecir si una persona lee la información nutrimental (variable dependiente P4, con los valores de SI y NO), el valor de interés es NO y la variable P7, no se considera.
      ii. Problema 2: Predecir la frecuencia con que una persona va a elegir un producto por la información de los empaques (variable dependiente P7, con los valores 1 a 5). No se considera P4.
   d. Leer los datos
      i. Crear 2 subconjuntos de datos, uno por problema.
      ii. Eliminar los renglones con NA y de las personas que no saben o no respondieron (código 9).
      iii. Sustituir en la variable dependiente, el valor numérico por su significado (código)
      iv. Reportar el número de renglones que quedó en cada uno.

Problema 1.

2. Entendimiento de los datos
   Análisis estadístico de los rasgos del vector de características (apoyarse de gráficos, de tamaño apropiado)
   a. Número de datos faltantes por columna.
   b. Análisis univariado.
   c. Análisis de correlación.
   d. Realizar una comparación de la distribución de cada una de las variables con respecto a la variable de interés.
   e. Otros análisis bivariados y/o multivariados.
   f. Escribir los hallazgos e información útil de este paso
3. Preparación de los datos
   Transformaciones e ingeniería de características, colocar criterio en que se basó cualquiera de las decisiones
   a. Omitir alguna característica, cambiar de numérica a texto o crear una nueva a partir de otras
   b. Sustitución (imputación)
      i. De los valores NA.
      ii. De las respuestas de No sabe/No responde, si se considera apropiado o no.

   c. En variables categóricas considerar
      i. Agrupar para no manejar demasiados valores.
      ii. Eliminar renglones con menos de 10 ocurrencias o agruparlos.
      iii. Crear variables binarias.
   d. En variables cuantitativas, sean discretas o continuas, considerar.
      i. Omitir el percentil 1 y el 99 de los datos, para disminuir outliers.
      ii. Crear categorías por ejemplo en edad si es bebe, niño, adolescente, adulto y adulto mayor.
   e. Realizar actividades de entendimiento de los datos, si se considera necesario ante los cambios.
4. Modelado
   a. Conocimiento previo
      i. Elegir 3 algoritmos y 2 predictores para generar las regiones de decisión.
      ii. Elegir 2 de los algoritmos, que se caractericen por ser lentos y usarlos de forma directa (sin resampling) con todos los atributos posibles, para medir el tiempo requerido para entrenar un modelo.
   b. Experimentos
      i. Especificar por algoritmo las variables a usar, así como si se van a escalar (normalizar o estandarizar) o no.
      ii. Explicar la estrategia de creación de CE, CV y CP, junto al método de remuestreo a utilizar.
      iii. Definir los valores de los hiperparámetros a probar.
      iv. Construir los modelos de todos los algoritmos bajo las condiciones establecidas.
5. Evaluación
   a. Probar todos los modelos con su hiperparámetro afinado en el CP.
   b. Hacer una tabla comparativa del rendimiento de los modelos.
   c. Plantear el problema de clasificación final.
   d. Ejecutar pruebas para aumentar el rendimiento del algoritmo con mejor desempeño, procurando ser más sensible.
6. Despliegue
   Implementar el código en R (crear el programa) con el mejor modelo y los siguientes pasos.
   a. Lee un archivo con datos llamado pruebaFinal.csv.
   b. Realiza las transformaciones necesarias.
   c. Divide en CE y CP (50 y 50) usando la semilla 2711.
   d. Crea el modelo con los hiperparámetros establecidos.
   e. Aplica el algoritmo al CP.
   f. Muestra las medidas de rendimiento obtenidas con caret de exactitud y el F1.

Problema 2.

7. Entendimiento de los datos
   a. Realizar una comparación de la distribución de cada una de las variables con respecto a la variable de interés.
   b. Escribir los hallazgos e información útil de este paso
8. Preparación de los datos: Aplicar los mismos criterios definidos en el Problema 1 (si aplica)
9. Modelado
   a. Experimentos
      i. Ocupar la estrategia de creación de CE, CV y CP, del Problema 1.
      ii. Decidir el método de resampling a utilizar.
      iii. Definir los valores de los hiperparámetros a probar.
      iv. Construir los modelos de todos los algoritmos aplicable evaluados por exactitud, bajo las condiciones establecidas.
10. Evaluación
   a. Probar todos los modelos con el CP.
   b. Hacer una tabla comparativa.
   c. Plantear el problema de clasificación final.
   d. Intentar mejorar el resultado, tal vez con otros hiperparámetros.

Proceso final

11. Interpretación de resultados
   a. Exponer todos los aprendizajes, que se consideran podrían ayudar a la persona de Negocio. Busca explicar e interpretar los resultados, apoyados con la información de todos los modelos.
   b. Conclusiones por integrante, dirigidas a la experiencia que te aportó el proceso realizado.
12. Referencias.