
Centro de Investigación en Matemáticas

Unidad Monterrey

Proyecto No. 1.

# Métodos multivariados de Análisis de Datos

## Análisis de nutrientes en pizzas.

Ricardo Cruz Sánchez
Rolando Corona Jiménez

# Índice

# 1. Introducción.

Aplicación de modelos de clasificación para obtener un filtro (personalizado) para correos electrónicos spam.

# 2. Análisis exploratorio.

## 2.1. Descripción del conjunto de datos.

El conjunto de datos proviene de una serie de correos electrónicos del personal de la empresa HP. Los correos etiquetados como spam fueron proporcionados por el administrador del servidor de correo de la empresa, mientras que los correos que no están etiquetados como spam corresponden a correos personales y de trabajo de George Forman, es por ello que palabras como *george* o código de área 650 son indicadores de no spam. La base de datos fue creada por Mark Hopkins, Erik Reeber, George Forman y Jaap Suermondt de Hewlett-Packard Labs.

En total se cuentan 4601 observaciones, de las cuales 1813 fueron etiquetadas como spam, que corresponde al 39.4 % del total. Las observaciones están representadas a través de un conjunto de 58 atributos: 57 variables continuas y una variable nominal de clase. Ninguno de los atributos presenta datos faltantes.

| | | |
|---|---|---|
| Spam | 1813 | 39.4 % |
| Non-Spam | 2788 | 60.6 % |

Tabla 1: Distribución de clases.

## 2.2. Diccionario de datos

Los 58 atributos se pueden agrupar en:

- 48 variables continuas con rango $[0, 100]$, de la forma word_freq_WORD, que indica el porcentaje de palabras en el correo electrónico que coinciden con WORD, es decir:

$$word\_freq\_WORD = 100 \times \frac{\#\text{apariones de WORD en el correo}}{\#\text{total de palabras en el correo}}$$

- 6 variables continuas

# 3. Métricas para la clasificación binaria.

## 3.1. Accuracy

## 3.2. F1-score

## 3.3. Selección de métricas para el problema del spam

# 4. Modelos de clasificación.

## 4.1. Regresión logística.

## 4.2. Máquinas de soporte vectorial.

## 4.3. Árboles de decisión.

## 4.4. Random Forest.

## 4.5. Modelos con reducción de dimensión y selección de variables

## 4.6. Comparativa con otros modelos

# 5. Análisis del dataset elron.

# 6. Conclusiones.

# Referencias


[1] Lorrie Faith Cranor and Brian A. LaMacchia. 1998. Spam!. Commun. ACM 41, 8 (August 1998), 74-83. DOI: https://doi.org/10.1145/280324.280336

[2] Emilio Ferrara. 2019. The history of digital spam. Commun. ACM 62, 8 (July 2019), 82-91. DOI: https://doi.org/10.1145/3299768

[3] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..