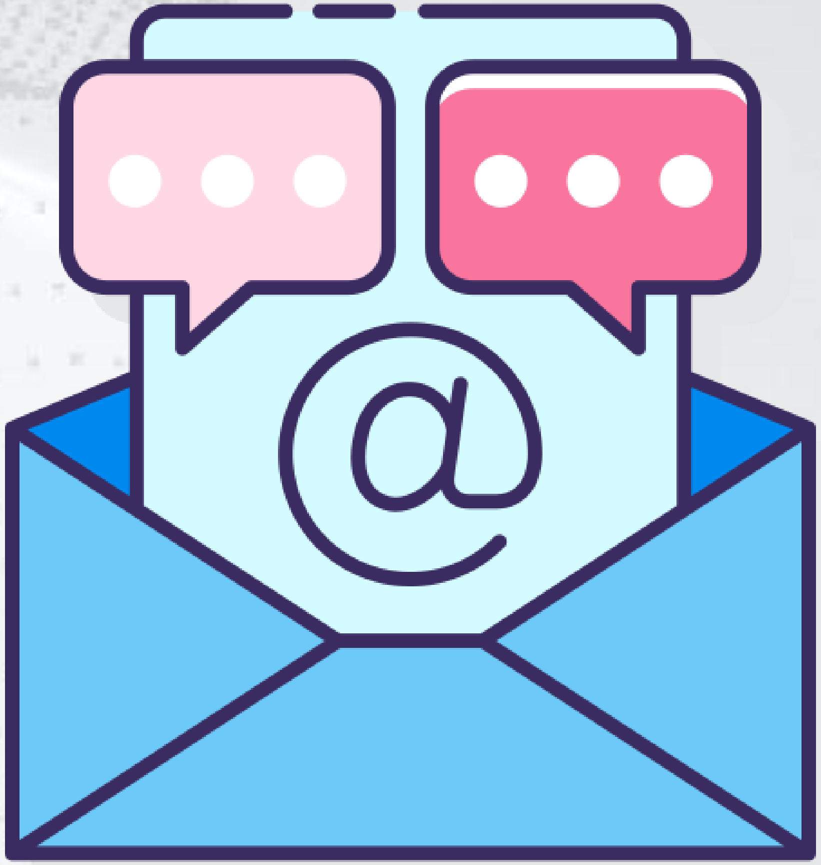
KIBERNUM
IT Academy

Módulo 5 – Aprendizaje de Máquina Supervisado

Actividad: Caso KNN Email Spam

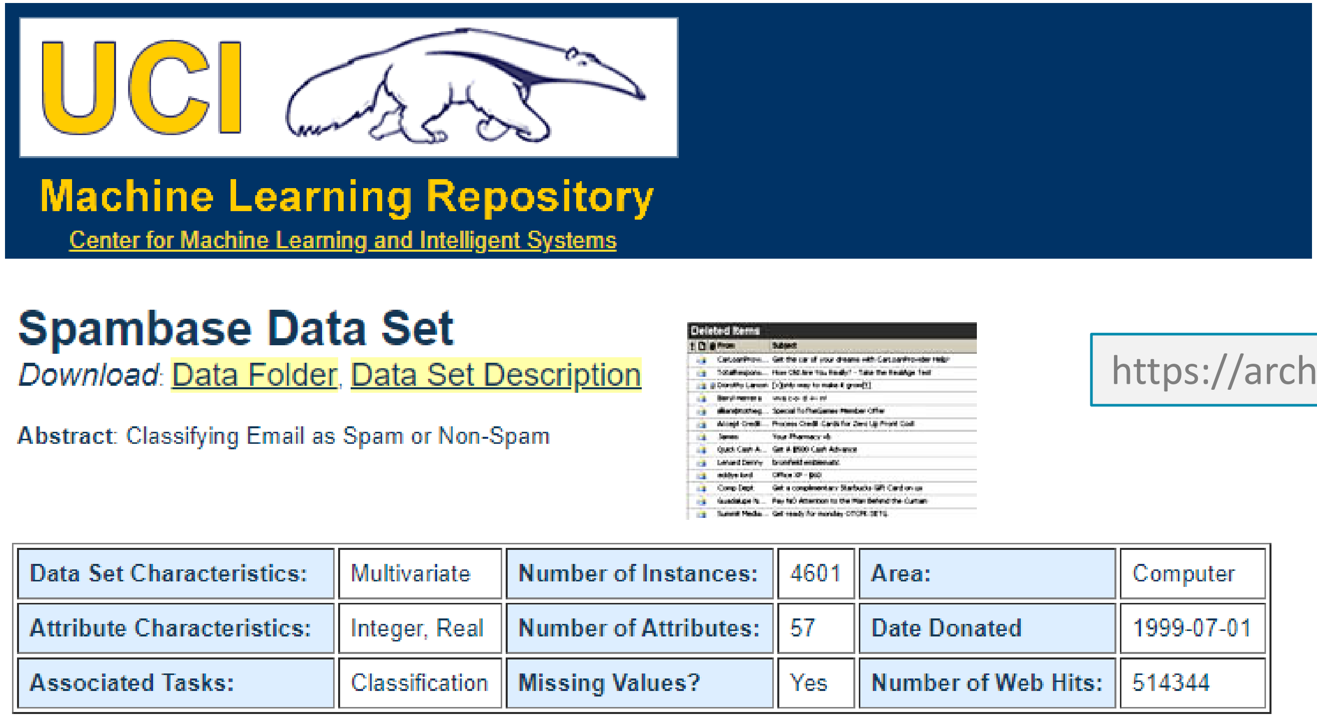
Especialización en Ciencia de Datos

Caso Email Spam



Caso Email Spam

El objetivo es entrenar un algoritmo de Clasificación KNN para detectar un correo spam. Para eso, utilizaremos el set de datos de spam.



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Spambase Data Set

Download: Data Folder, Data Set Description

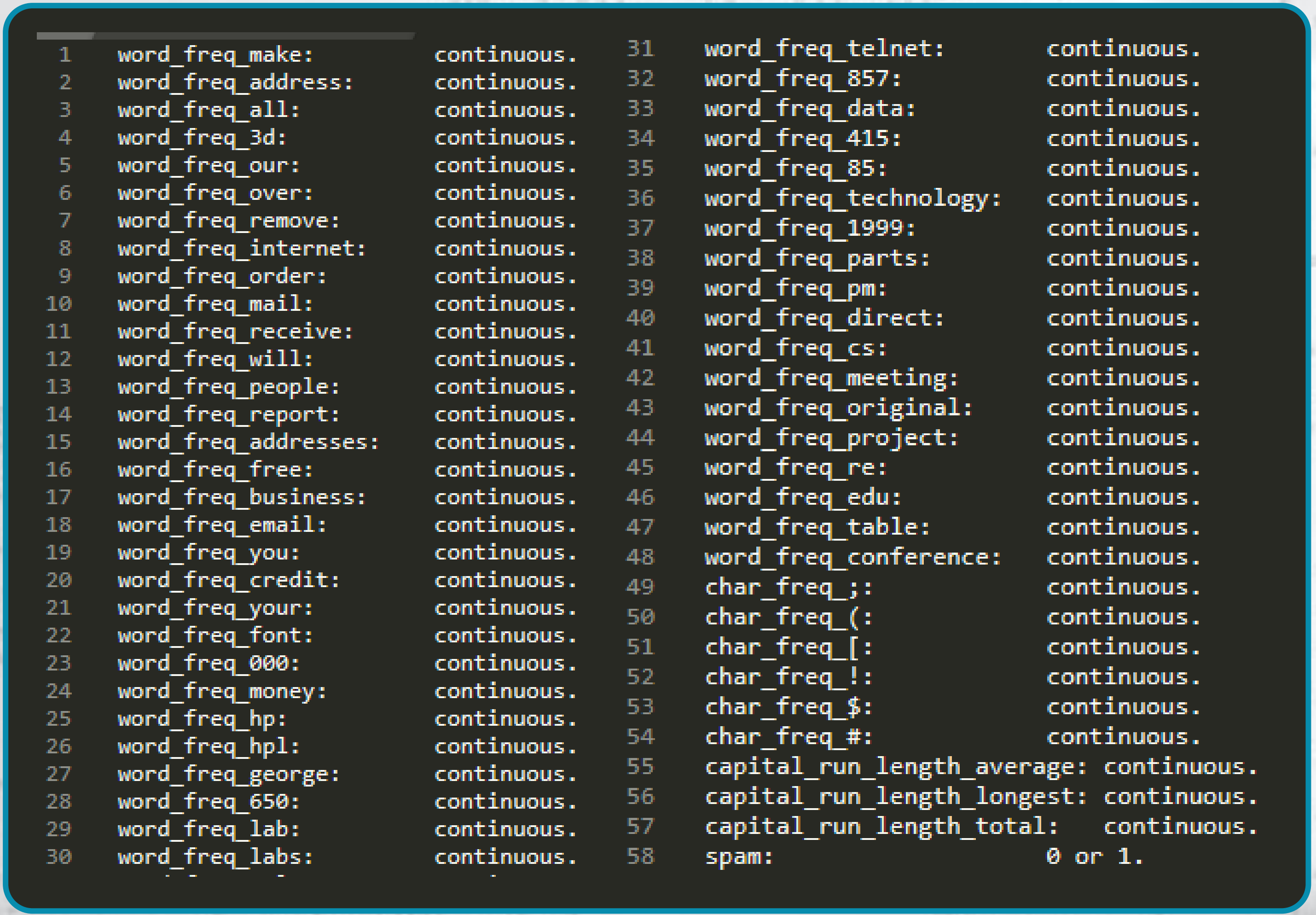
Abstract: Classifying Email as Spam or Non-Spam

Data Set Characteristics:	Multivariate	**Number of Instances:**	4601	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	57	**Date Donated**	1999-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	514344

https://archive.ics.uci.edu/ml/datasets/spambase

Caso UN Comtrade

Los 58 valores corresponden a los siguientes campos:



```
1   word_freq_make:          continuous.
2   word_freq_address:       continuous.
3   word_freq_all:           continuous.
4   word_freq_3d:            continuous.
5   word_freq_our:           continuous.
6   word_freq_over:          continuous.
7   word_freq_remove:        continuous.
8   word_freq_internet:      continuous.
9   word_freq_order:         continuous.
10  word_freq_mail:          continuous.
11  word_freq_receive:       continuous.
12  word_freq_will:          continuous.
13  word_freq_people:        continuous.
14  word_freq_report:        continuous.
15  word_freq_addresses:     continuous.
16  word_freq_free:          continuous.
17  word_freq_business:      continuous.
18  word_freq_email:         continuous.
19  word_freq_you:           continuous.
20  word_freq_credit:        continuous.
21  word_freq_your:          continuous.
22  word_freq_font:          continuous.
23  word_freq_000:           continuous.
24  word_freq_money:         continuous.
25  word_freq_hp:            continuous.
26  word_freq_hpl:           continuous.
27  word_freq_george:        continuous.
28  word_freq_650:           continuous.
29  word_freq_lab:           continuous.
30  word_freq_labs:          continuous.
```

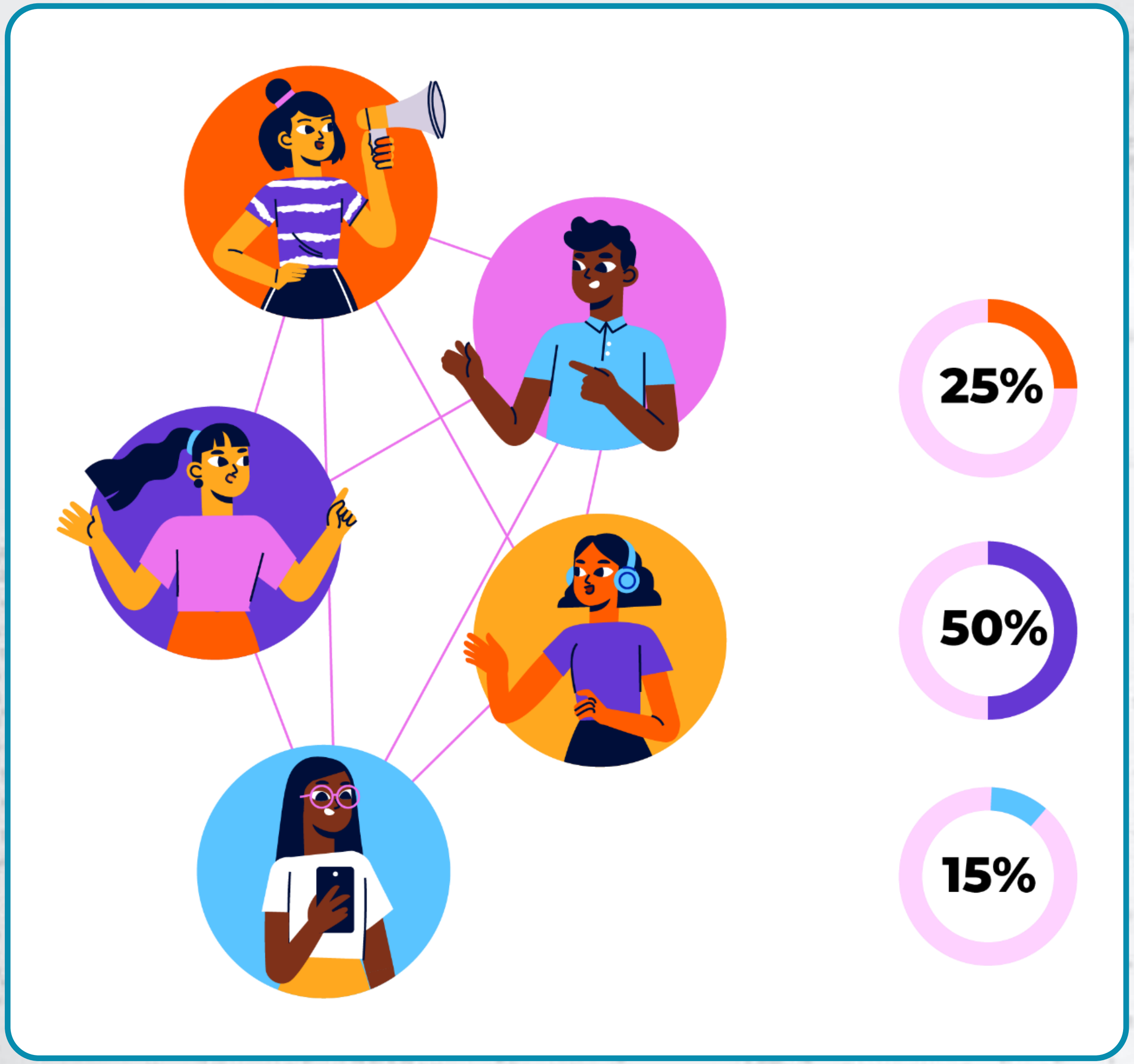
```
31  word_freq_telnet:            continuous.
32  word_freq_857:               continuous.
33  word_freq_data:              continuous.
34  word_freq_415:               continuous.
35  word_freq_85:                continuous.
36  word_freq_technology:        continuous.
37  word_freq_1999:              continuous.
38  word_freq_parts:             continuous.
39  word_freq_pm:                continuous.
40  word_freq_direct:            continuous.
41  word_freq_cs:                continuous.
42  word_freq_meeting:           continuous.
43  word_freq_original:          continuous.
44  word_freq_project:           continuous.
45  word_freq_re:                continuous.
46  word_freq_edu:               continuous.
47  word_freq_table:             continuous.
48  word_freq_conference:        continuous.
49  char_freq_;:                 continuous.
50  char_freq_(:                 continuous.
51  char_freq_[:                 continuous.
52  char_freq_!:                 continuous.
53  char_freq_$:                 continuous.
54  char_freq_#:                 continuous.
55  capital_run_length_average: continuous.
56  capital_run_length_longest: continuous.
57  capital_run_length_total:   continuous.
58  spam:                        0 or 1.
```

Caso Email Spam

Construya un notebook jupyter ordenado, documentado y reproducible, en donde entrene un modelo predictivo de clasificación utilizando el algoritmo KNN.

Usted tendrá que realizar lo siguiente:

1. Accuracy del algoritmo.
2. Matriz de Confusión.
3. Valor de K seleccionado.
4. Subir este notebook en su carpeta de trabajos.



¡Éxito!