Máster Universitario en Ingeniería Informática

Tratamiento Inteligente de Datos

PRÁCTICA 1: Preprocesamiento

Accidentes

UNIVERSIDAD DE GRANADA

Carlos Santiago Sánchez Muñoz

Grupo de prácticas 1 - Jueves

Email: carlossamu7@correo.ugr.es

4 de noviembre de 2020

Índice

1. Discretización

El primer problema a abordar es un problema de clasificación. La base de datos es a usar es *Optical Recognition of Handwritten Digits* que contiene imágenes de dígitos del sistema de numeración arábigo. El objetivo consiste en aprender de esta base de datos para poder clasificar otras imágenes con dígitos manuscritos.

```
""" Lectura de datos. Devuelve dos df, uno de ellos con las columnas imputadas.
- mini (op): indica si leer 'accidentes_mini'. Por defecto 'True'.
"""
def read_data(mini=True):
    if(mini):
        df = pd.read_excel(r'accidentes_mini.xls', sheet_name='datos')
    else:
        df = pd.read_excel(r'accidentes.xls', sheet_name='datos')

    df_I = df.copy()
    df = df.drop(columns = ['WKDY_I', 'HOUR_I', 'MANCOL_I', 'RELJCT_I', 'ALIGN_I',
                'PROFIL_I', 'SURCON_I', 'TRFCON_I', 'SPDLIM_H',
                'LGTCON_I', 'WEATHR_I', 'ALCHL_I'])
    df_I = df_I.drop(columns = ['WEEKDAY', 'HOUR', 'MAN_COL', 'REL_JCT', 'ALIGN',
                'PROFILE', 'SUR_COND', 'TRAF_CON', 'SPD_LIM',
                'LGHT_CON', 'WEATHER', 'ALCOHOL'])
    if(IMPRIME_INFO):
        print()
        print(df.info())
        print()
        print(df_I.info())
    return df, df_I
```

2. Valores perdidos

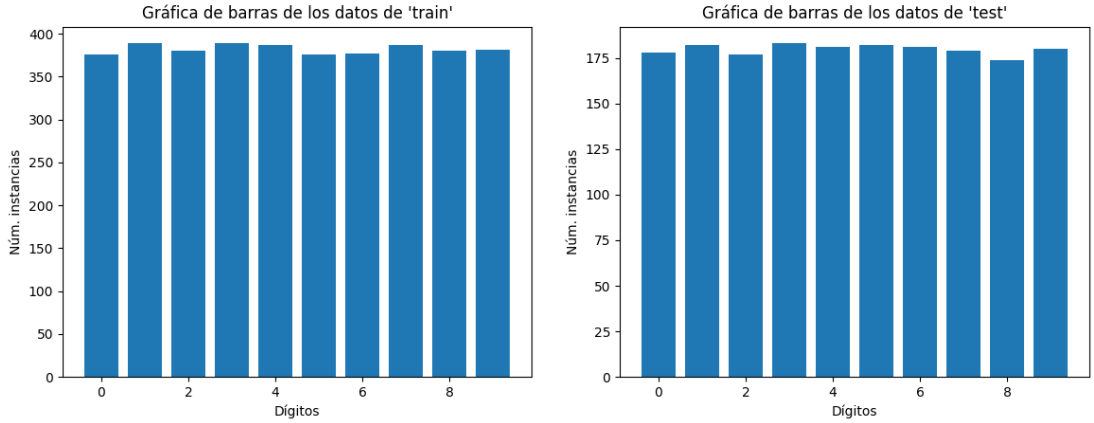
3. Selección de características

4. Selección de instancias

Tabla 1: Reparto de los dígitos en 'train' y 'test'

Díg.	**Instancias 'train'**	**Porcentaje 'train' (%)**	**Instancias 'test'**	**Porcentaje 'test' (%)**
0	376	9.84	178	9.91
1	389	10.18	182	10.13
2	380	9.94	177	9.85
3	389	10.18	183	10.18
4	387	10.12	181	10.07
5	376	9.84	182	10.13
6	377	9.86	181	10.07
7	387	10.12	179	9.96
8	380	9.94	174	9.68
9	382	9.99	180	10.02

En unas gráficas de barras lo podemos apreciar mejor:



(a) Train (b) Test

Imagen 1: Gráficas de barras de los datos