# Pump it Up. Equipo CharlaTED

Ignacio Aguilera Martos (nacheteam, KNN), Luis Balderas Ruiz (luisbalru, RIP-PER), Francisco Luque Sánchez (¡usuario¿, ¡algoritmo¿), Iván Sevillano García ¡usuario¿, SVM

18 de febrero de 2020

Preprocesamiento y Clasificación

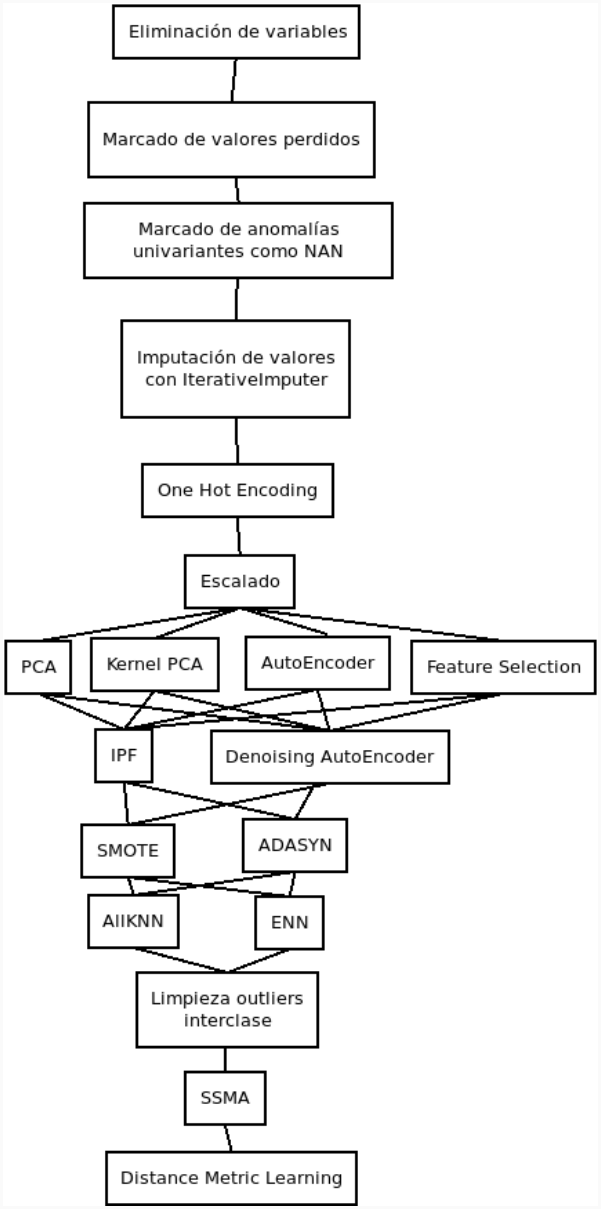
# Contenidos

# SVM

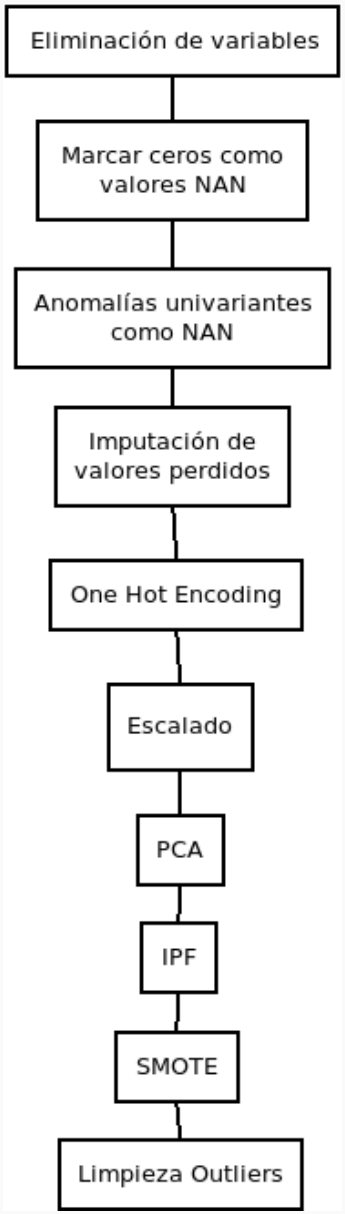
# RIPPER

# KNN

# Pipeline empleado



# Pipeline con mejor resultado



# Explicación de las técnicas

# Explicación de las técnicas

**Eliminación de variables**

Elimino las variables wpt_name, subvillage, scheme_name, funder, installer, ward, amount_tsh y num_private.

# Explicación de las técnicas

## Eliminación de variables

Elimino las variables wpt_name, subvillage, scheme_name, funder, installer, ward, amount_tsh y num_private.

## Marcado de anomalías como valores perdidos

En cada columna se calcula la media y la desviación típica. Aquellos datos que se salgan del intervalo $[media - 5std, media + 5std]$ se marcan como NAN.

# Explicación de las técnicas

## Eliminación de variables

Elimino las variables wpt_name, subvillage, scheme_name, funder, installer, ward, amount_tsh y num_private.

## Marcado de anomalías como valores perdidos

En cada columna se calcula la media y la desviación típica. Aquellos datos que se salgan del intervalo $[media - 5std, media + 5std]$ se marcan como NAN.

## Imputación iterativa

Empleamos una imputación iterativa sobre los valores perdidos.

# Explicación de las técnicas

## Eliminación de variables

Elimino las variables wpt_name, subvillage, scheme_name, funder, installer, ward, amount_tsh y num_private.

## Marcado de anomalías como valores perdidos

En cada columna se calcula la media y la desviación típica. Aquellos datos que se salgan del intervalo $[media - 5std, media + 5std]$ se marcan como NAN.

## Imputación iterativa

Empleamos una imputación iterativa sobre los valores perdidos.

## PCA

Aplicamos PCA pero sólo sobre las columnas categóricas. El objetivo es explicar las variables categóricas mejor que en su codificación original. Reducimos a 44 variables todas las categóricas.

# Explicación de las técnicas

# Explicación de las técnicas

**IPF**

Ejecutamos un IPF para limpiar el ruido con 4 iteraciones.

# Explicación de las técnicas

## IPF

Ejecutamos un IPF para limpiar el ruido con 4 iteraciones.

## SMOTE

Hacemos un oversampling de las clases "functional needs repair" y "non functional" a 7500 y 22000 con respecto a 23500 de la clase "functional" con $k = 7$.

# Explicación de las técnicas

## IPF

Ejecutamos un IPF para limpiar el ruido con 4 iteraciones.

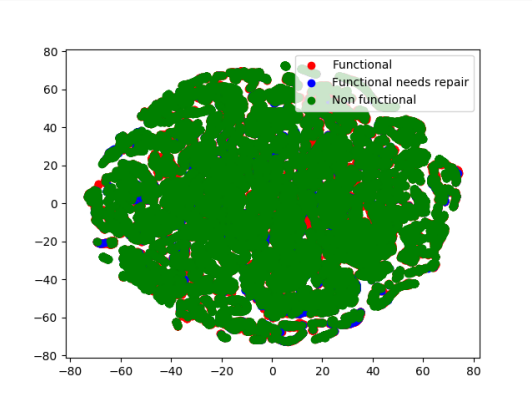
## SMOTE

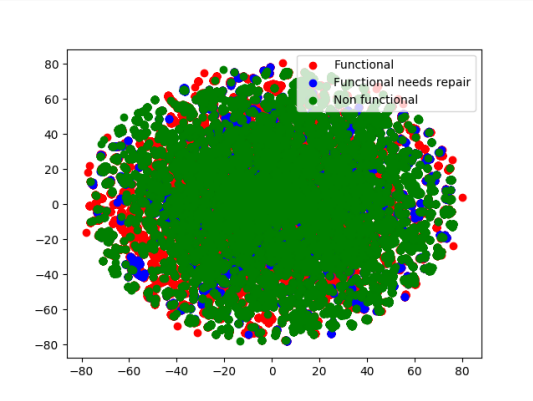
Hacemos un oversampling de las clases "functional needs repair" y "non functional" a 7500 y 22000 con respecto a 23500 de la clase "functional" con $k = 7$.

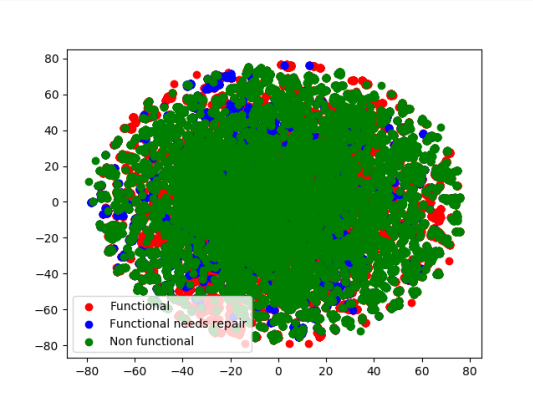
## Limpieza Outliers

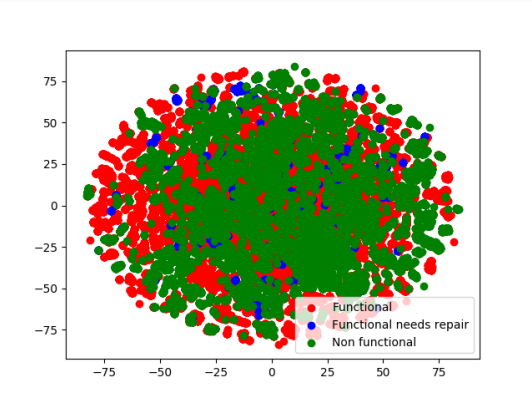
Hacemos una limpieza de anomalías por cada clase eliminando el 1 % más anómalo según KNN con $k = 7$ y la métrica de la mayor distancia.

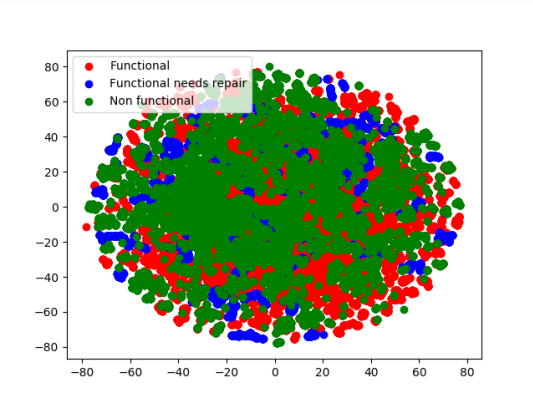
# Visualización de las técnicas

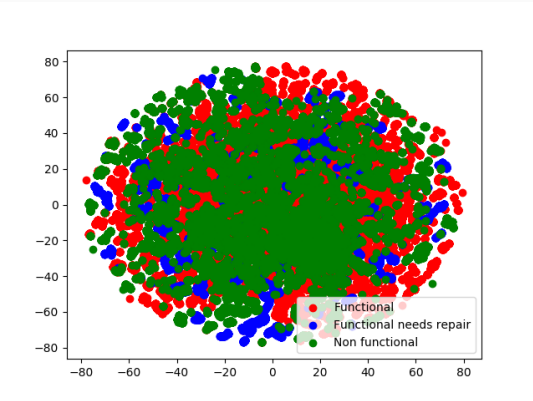




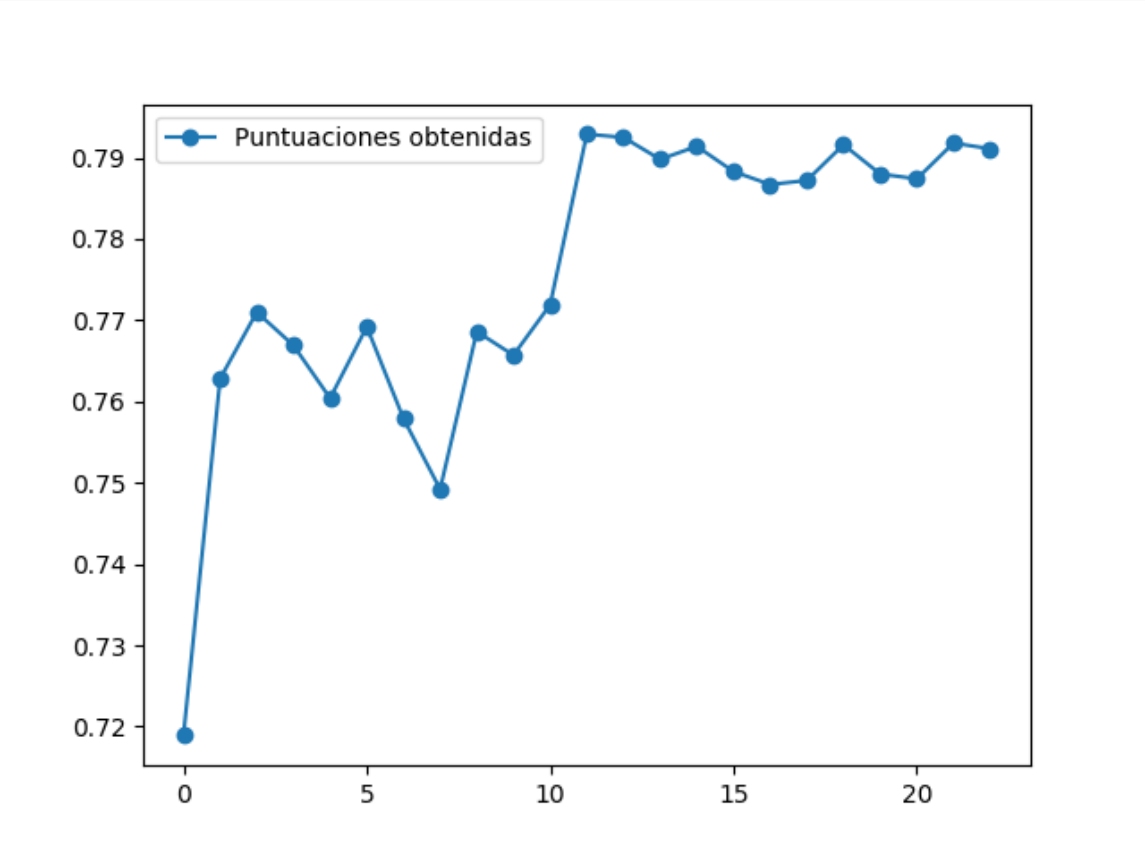








# Posición en DrivenData




Puntuación final obtenida: 79.29 %

Ranking final: 1729

Número de subidas: 23

# J48

# Referencias i

# ¿Preguntas?