

Pump it Up. Equipo CharlaTED

Ignacio Aguilera Martos (nacheteam, KNN), Luis Balderas Ruiz (luisbalru, RIP-PER), Francisco Luque Sánchez (jusuuarioĭ, jalgoritmoĭ), Iván Sevillano García (jusuuarioĭ, SVM)

17 de febrero de 2020

Preprocesamiento y Clasificación

1. SVM
2. SVM
3. RIPPER
4. KNN
5. J48

SVM

SVM

- Eliminación de variables.
 - *region, recorded_by, num_private,...*
 - Variables categóricas con más de 100 valores.

Preprocesamiento para SVM

- Eliminación de variables.
 - *region*, *recorded_by*, *num_private*,...
 - Variables categóricas con más de 100 valores.
- Detección de valores perdidos(NA).
 - *population* = 0??
 - *construction_year* = 0??
 - Valores vacíos.

Preprocesamiento para SVM

- Eliminación de variables.
 - *region*, *recorded_by*, *num_private*,...
 - Variables categóricas con más de 100 valores.
- Detección de valores perdidos(NA).
 - *population* = 0??
 - *construction_year* = 0??
 - Valores vacíos.
- Cambiar tipos de dato.
 - *region_code*/*district_code* a factor.
 - *date* a numérico.

IPF(Iterative partitioning filter)

- Split the current training dataset E into n equal sized subsets.
- Build a classifier with the C4.5 algorithm over each of these n subsets and use them to evaluate the whole current training dataset E .
- Add to A the noisy examples identified in E using a voting scheme (consensus or majority).
- Remove the noisy examples.

Sáez, J. A., Luengo, J., Stefanowski, J., Herrera, F. (2015).

SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering.

Information Sciences, 291, 184-203.

Imputación de valores perdidos

- Amelia para valores perdidos numéricos(método iterativo de imputación).
- Imputación mediante KNN de valores categóricos(modas).

Duminicación de variables

- Para cada valor las variables categórica, creamos una variable que valdrá 1 si la instancia tiene este valor.
- Eliminamos una de las variables dumificadas ya que no aporta más información.

PCA para variables dumificadas. Selección de subespacios relevantes

- Normalizamos las variables para aplicar PCA.
- Aplicamos PCA y nos quedamos con las variables cuya desviación típica sea mayor de un umbral(0.00001).
- Pasamos de 208 variables a 132.

Aplicación de SMOTE para balancear clases

El método SMOTE genera instancias de la clase minoritaria, sobrerrepresentandola. Mejora la clasificación de instancias minoritarias pero no la tasa de acierto.

Flujo de información

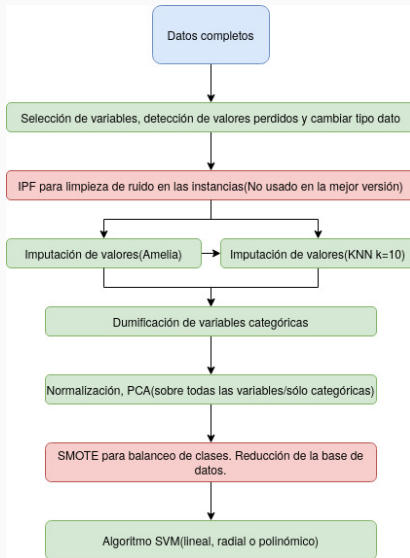


Figura 1: Verde: Utilizado en el mejor modelo

Puntuación a lo largo del tiempo de SVM

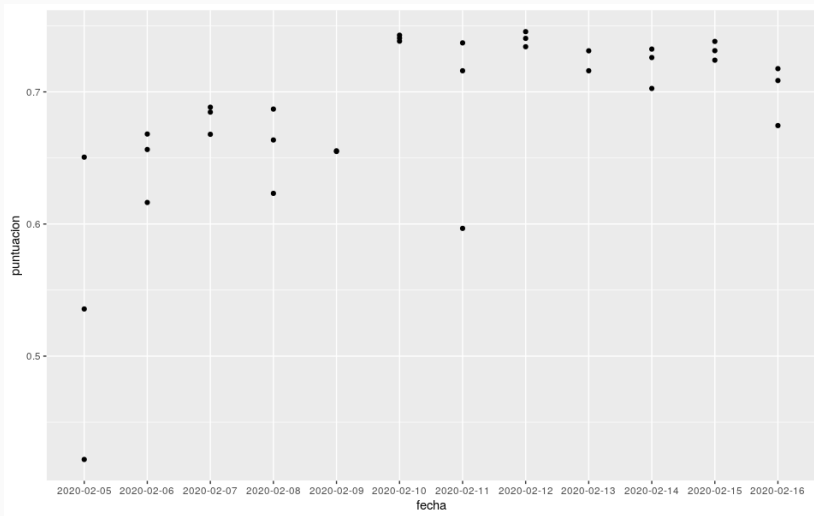


Figura 2: Máxima puntuación: 74.52 % el 12 de febrero

RIPPER

Algoritmos y técnicas utilizados

- **Ingeniería de características Selección y creación de características**
 - Selección de variables semánticamente representativas.
 - LasVegas Wrapper.
 - Imputación de valores perdidos con media o mediana.
 - Creación de características.

Algoritmos y técnicas utilizados

- **Ingeniería de características Selección y creación de características**
 - Selección de variables semánticamente representativas.
 - LasVegas Wrapper.
 - Imputación de valores perdidos con media o mediana.
 - Creación de características.
- **Técnicas basadas en instancias. Selección, eliminación de ruido, sobremuestreo**
 - ENN.
 - IPF.
 - SMOTE. Oversampling
 - Random Undersampling.

Algoritmos y técnicas utilizados

- **Ingeniería de características Selección y creación de características**
 - Selección de variables semánticamente representativas.
 - LasVegas Wrapper.
 - Imputación de valores perdidos con media o mediana.
 - Creación de características.
- **Técnicas basadas en instancias. Selección, eliminación de ruido, sobremuestreo**
 - ENN.
 - IPF.
 - SMOTE. Oversampling
 - Random Undersampling.
- **Ajuste de hiperparámetros.**
 - Gridsearch de parámetros F (número de folds), N (mínimo peso de instancias) y O (número de ejecución para optimizar).
 - 5-CV.

contenidos...

Resultados

- 27 subidas. Mejor resultado: 0.7869.
- Mejor posición: 1752. Posición actual (17/02): 1834

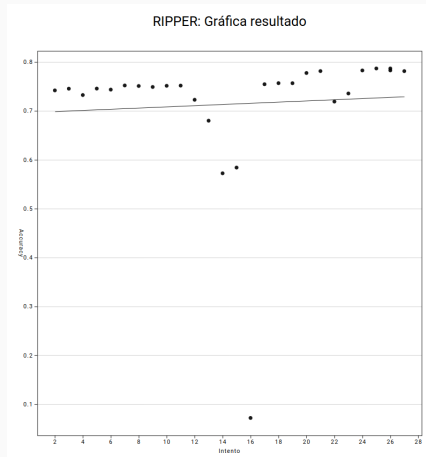


Figura 3: Gráfica de resultados

KNN

J48

¿Preguntas?