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

SVM

Preprocesamiento para 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 vacios.

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 vacios.
- Cambiar tipos de dato.
  - *region_code*/*district_code* a factor.
  - *date* a numérico.

Limpieza de ruido mediante IPF

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(moda).

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, sobrerepresentandola. Mejora la clasificación de instancias minoritarias pero no la tasa de acierto.

Flujo de información

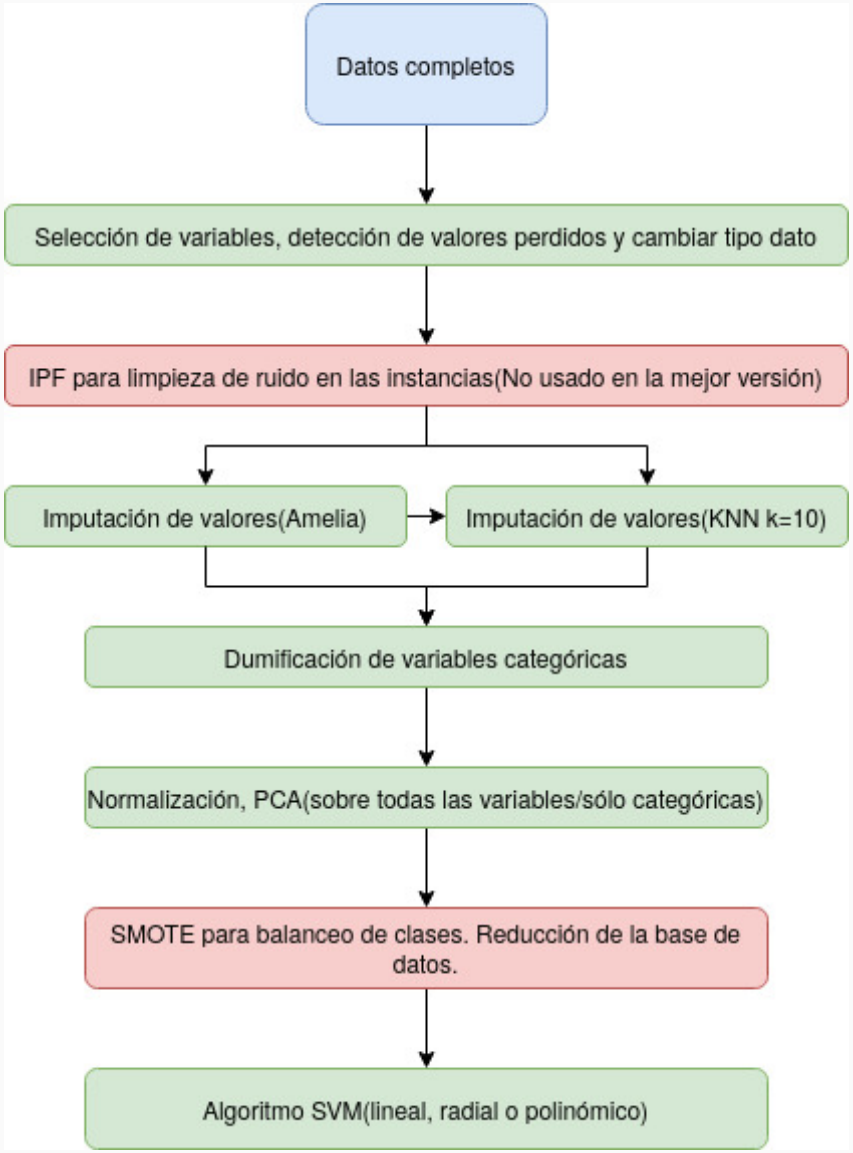

Figura 1: Verde: Utilizado en el mejor modelo

Puntuación a lo largo del tiempo de SVM

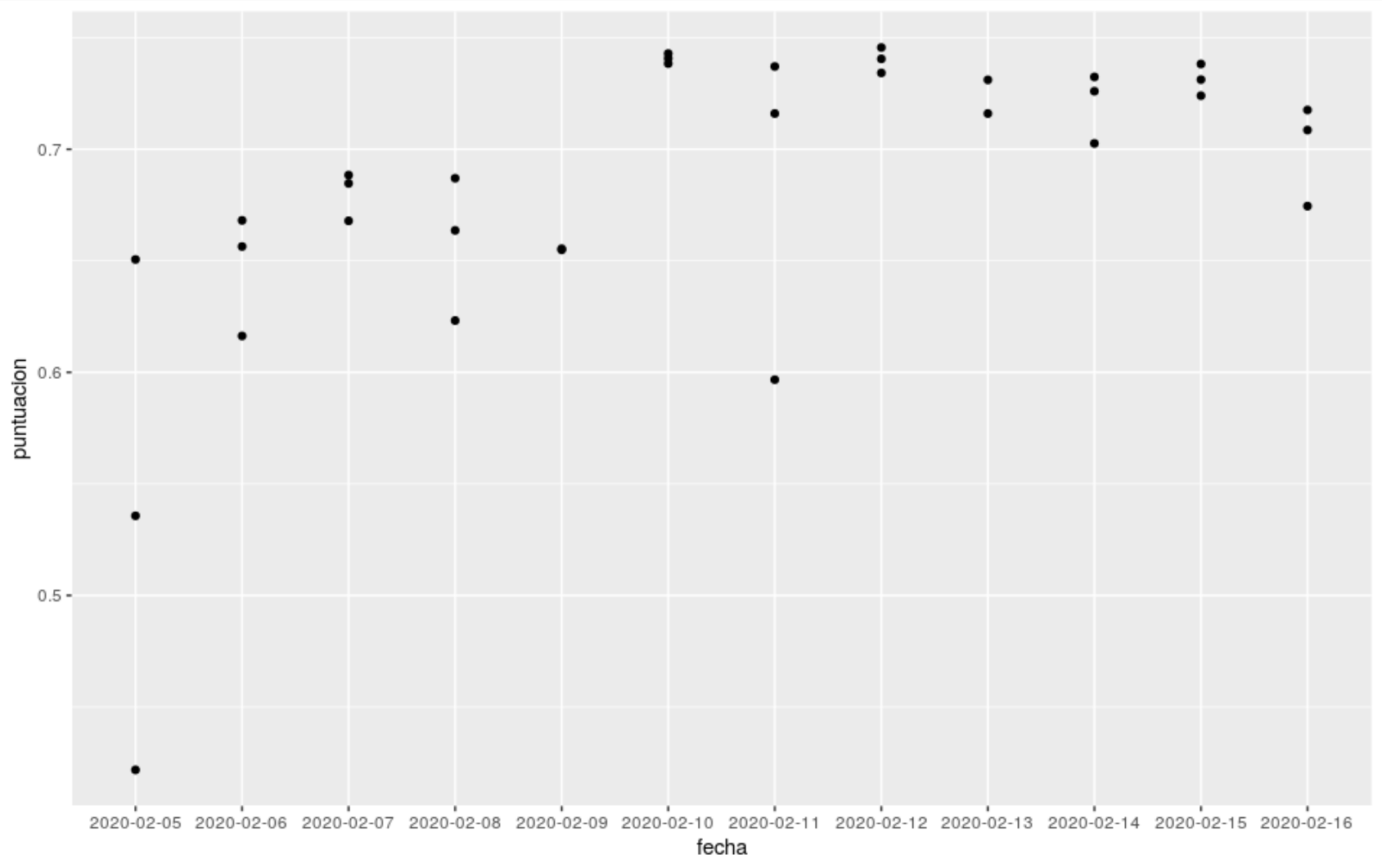

Figura 2: Máxima puntuación: 74.52 % el 12 de febrero

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.

Malos resultados

LasVegas Wrapper

- Implementación en R (FSinR) que no consigue terminar una ejecución completa.

Malos resultados

LasVegas Wrapper

- Implementación en R (FSinR) que no consigue terminar una ejecución completa.

ENN, IPF

- Limpieza de ruido empeora los resultados tanto en validación cruzada como test. Para ciertas configuraciones, incluso hacen desaparecer la clase minoritaria.

Malos resultados

SMOTE, Oversampling. Random Undersampling

- Necesidad de 'numerizar' los datos. Creación de variables dummies.

Malos resultados

SMOTE, Oversampling. Random Undersampling

- Necesidad de 'numerizar' los datos. Creación de variables dummies.
- Las técnicas de oversampling generan datos demasiado artificiales. En general, tienen accuracy cercano al 58 %.

Imputación de valores perdidos o mal escritos

Malos resultados

SMOTE, Oversampling. Random Undersampling

- Necesidad de 'numerizar' los datos. Creación de variables dummies.
- Las técnicas de oversampling generan datos demasiado artificiales. En general, tienen accuracy cercano al 58 %.

Imputación de valores perdidos o mal escritos

- *Funder:* factor de 2141 niveles.
- *Installer:* factor de 2411 niveles.
- Muchos de esos niveles son resultado de escribir mal o de distintas formas la misma palabra.
- Reducción de los niveles hasta aproximadamente 500 corrigiendo los nombres, con peor precisión como resultado.

Buenos resultados

Selección de características: Muchas columnas parecen inútiles

- Elimino las variables *wpt-name*, *subvillage*, *ward*, *recorded-by*, *scheme-name*, *num-private*,*region-code*, *quantity-group*, *source-type*, *waterpoint-type-group*, *payment-type* y *extraction-type-group*
- Imputo valores perdidos en *funder*, *installer*, *permit*, *scheme-management*, *public-meeting*, *gps-height*, *extraction-type.*
- Imputación de la variable *construction-year* con los resultados de la validación cruzada en entrenamiento del mejor algoritmo hasta el momento (KNN).

Imagen del conjunto tras el preprocesamiento (TSNE)

Mejor modelo

- Ripper (JRip de RWeka) con variables *latitude, longitude, date-recorded,basin,lga,funder,population,construction-year, gps-height,public-meeting,scheme-name,permit,extraction-type-class,management, management-group, payment, quality-group, quantity, source, source-type,source-class y waterpoint-type.*
- Hiperparámetros $F = 2, N = 3, O = 29$.

Resultados

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

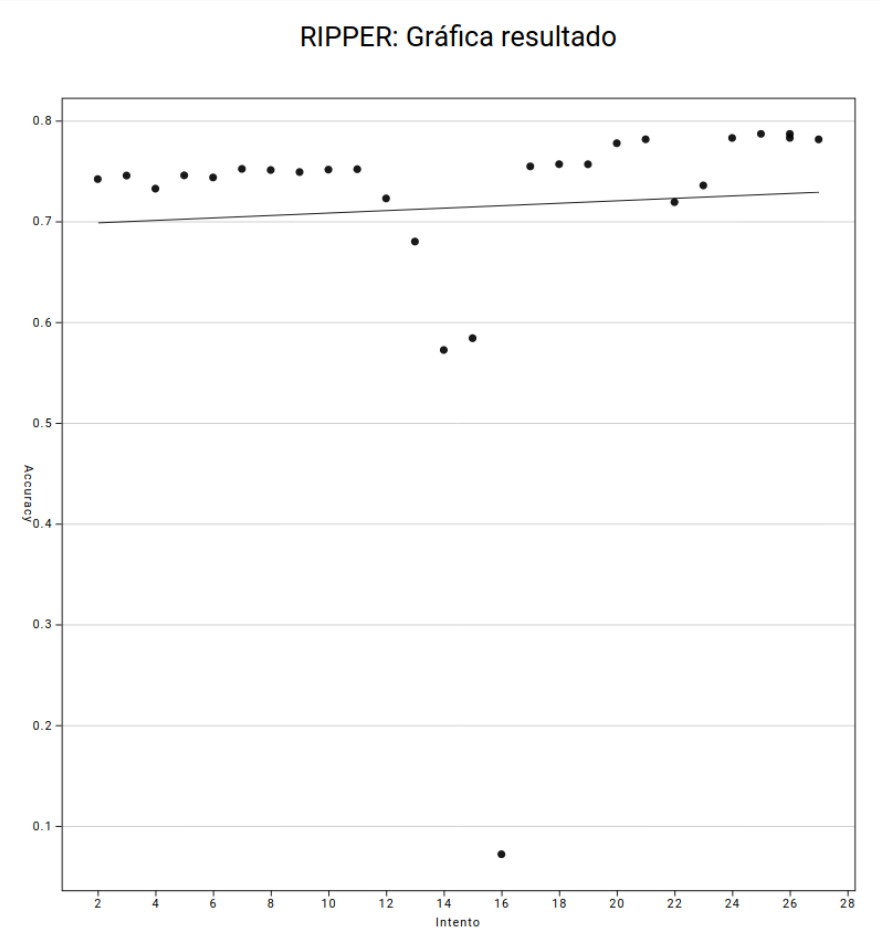

Figura 3: Gráfica de resultados

KNN

J48

¿Preguntas?