Clasificación de granizo en superficie usando técnicas de minería de datos y datos de radar meteorológico

Yanina Bellini Saibene[1, 2], Martín Volpacchio[2]

[1], EEA INTA Anguil
Ruta Nac. Nº 5 Km 580, (6326) Anguil, La Pampa, Argentina
{bellini.yanina}@inta.gob.ar
[2] Universidad Austral (Maestría en Data Mining),

Resumen. El granizo es capaz de infringir cuantiosos daños y su estudio es útil para los servicios meteorológicos, la industria de los seguros y el sector agropecuario. Debido a la reducida ocurrencia espacial y temporal de los eventos de granizo, determinar su localización y frecuencia es difícil y costoso. Se realizó un estudio de los datos de un radar polarimétrico de banda C (INTA Anguil, La Pampa) utilizando 14 tormentas de granizo y alrededor de 1400 casos en un radio de 240 km desde Enero de 2011 a Diciembre de 2012. Se utilizaron cinco técnicas de minería de datos para calcular la probabilidad de granizo en superficie obteniendo una *accuracy* por encima del 86% con bajos valores (<21%) de falsas alarmas para el mejor modelo. Estos métodos probaron ser herramientas útiles para la clasificación de granizo con datos de radar.

1 Introducción

El granizo es un fenómeno meteorológico capaz de infligir cuantiosos daños [1],[2],[3]; se considera un riesgo agroclimático y su estudio es útil para los servicios meteorológicos, la industria de los seguros y la comunidad agropecuaria [2]. Es poco frecuente, con una reducida extensión espacial y temporal y una variabilidad que supera a la de los otros fenómenos meteorológicos [3],[4],[5],[6] por lo tanto detectar su ocurrencia en superficie es una tarea difícil y costosa [3],[4],[5],[6]. Ante esta situación los radares meteorológicos son una alternativa a las redes terrestres de mediciones, porque abarcan una gran superficie y disponen de una única resolución en tiempo y espacio [2]. Existen estudios que exploran la relación de las variables medidas por los radares con el granizo. La mayoría de estos trabajos que utilizan técnicas de minería de datos (MD), realiza aprendizaje supervisado, por lo que necesitan un conjunto de datos previamente etiquetado para aprender y que posteriormente permita identificar la presencia de un hidrometeoro, en este caso granizo, en una nueva tormenta. Por ejemplo, para clasificar celdas de tormentas severas[1] hay estudios con *Redes Neuronales* (NN) [7], *Algoritmos Genéticos* (GA) [8], *Maquinas de Vectores Soporte* (SVM)

[1] Que pueden generar granizo

[9] y *Radial Basis Function* (RBF) [9]. También se usaron *Árboles de Decisión* (DT) [10],[11], *Regresiones* (R) [12] [13] [14], *Naive Bayes (NB)* [11] y *Redes Neuronales Bayesianas* (BNN) [15] para determinar la probabilidad de granizo severo, tamaño del granizo o tipo de tormenta. En Argentina los trabajos se concentran en la provincia de Mendoza y se utiliza *Regresión Logística* (RL) [3],[5]. El Instituto Nacional de Tecnología Agropecuaria (INTA) cuenta con un radar meteorológico de doble polarización en la provincia de La Pampa, cuya frecuencia de granizo es una de las más alta en la Pampa Húmeda [16]. El objetivo de este trabajo es generar un modelo de clasificación de ocurrencia de granizo en superficie utilizando técnicas de MD y datos polarimétricos derivados del radar meteorológico de INTA en La Pampa.

2 Datos

Para analizar el granizo se necesitan datos con alta resolución espacial y temporal debido a la pequeña escala y corta duración de este fenómeno [1]. Estos datos se obtuvieron de diversas fuentes que se integraron y organizaron en una base de datos con dos tipos de información: 1) los datos de campo (detallan la caída de granizo, después de una tormenta) y 2) los datos registrados por el radar.

2.1 Datos de campo

Los casos etiquetados, se recolectaron de fuentes usadas en los antecedentes:

a) Reportes de compañías de seguros agrícolas: SanCor, La Segunda y La Dulce ([14],[17],[18],[19],[20],[21]). Cuenta con verificación visual in situ de un perito.

b) Reportes en medios de comunicación y redes sociales ([13],[14],[20],[21],[22], [23]).

c) Redes de informantes ([14],[20],[21],[22],[23]): c.1)Red del Servicio Meteorológico Nacional (SMN) http://www.smn.gov.ar/?mod=voluntarios&id=1; c.2)Red de pluviómetros de la Policía de La Pampa http://www.policia.lapampa.gov.ar/lluvias.php; c.3)La red termo pluviométrica de la Red de Información Agropecuaria Nacional (RIAN) http://rian.inta.gov.ar/agua/informes.aspx; c.4)El Sistema Integrado de Información Agropecuaria (SIIA) http://www.siia.gov.ar/. Todas estas fuentes tienen observación in situ.

d)Informantes calificados: técnicos INTA, profesionales-asesores agropecuarios y productores locales ([3],[12],[16],[20],[21],[22],[23]). Hay observación in-situ.

e)Recorridas a campo posteriores o durante una tormenta ([21],[22],[24]): además de aprovechar las recorridas mensuales de la RIAN se realizaron salidas después de las tormentas del 10-12-2012 y el 24-12-2012.

Todas las fuentes proveen: la fecha de la tormenta, latitud y longitud del lugar y la etiqueta que indica si cayó o no cayó granizo. Se obtuvieron 1.419 lotes (1.077 negativos y 342 positivos) correspondientes a 14 fechas con tormentas del período primavero-estival del 2011 y 2012.

2.2 Datos del Radar Meteorológico

El radar ubicado en Anguil (La Pampa), opera en banda C y es de doble polarización. La antena permite un giro en sentido horizontal de 360º y este radar está configurado para elevarse en ángulo vertical 12 veces, entre 0,5º de base y 15,1º de tope, para rangos de 120, 240 y 480 km con una resolución espacial de 1 km^2. La frecuencia de este escaneo completo está programada cada 10 minutos, totalizando 144 adquisiciones diarias registrando las variables: factor de reflectividad (Z), reflectividad diferencial (Z_{DR}), coeficiente de correlación polarimétrica (Rho_{HV}), desplazamiento de fase diferencial (Phi_{DP}), desplazamiento de fase diferencial específica (K_{DP}), velocidad radial (V) y anchura del espectro (W). El día del radar se extienden de las 00:00 hs a las 23:50 hs [21].

Para este trabajo se procesaron Z, Z_{DR} y Rho_{HV} en el rango de 240 km, para la primera elevación (0,5º) integrando un valor por día; de esta manera los datos del radar y los de campo tienen la misma escala temporal y se pueden aparear. Para cada pixel de 1 km^2 se calculó el valor máximo, mínimo, promedio y total de las 144 tomas diarias correspondientes a los 14 días sobre los cuales se recolectaron los datos de campo. Estas variables se calcularon con la primera elevación porque al ser la más cercana a la superficie es la que mejor representa lo que puede precipitar a nivel del suelo [2],[17],[18],[22],[24] (Fig. 1). Los cálculos se realizaron con *GIC*, parte del software INTA-Radar[2] desarrollado en Python 2.7 que usa *numpy* para cálculos matriciales y *gdalg* para transformar las matrices en imágenes GeoTIFF [21]. El listado final de variables calculadas se presenta en la tabla 1.

Para conformar el dataset final, a las variables listadas en la tabla 1 se le agregó la variable etiquetada *Hail* que contiene el valor 1 en caso de caída de granizo y el valor 0 en caso que no haya caído granizo. Para realizar esta unión de los datos del radar con la de verdad de campo, también se usó el software INTA-Radar, pasando como parámetros: fecha, latitud y longitud del lugar etiquetado. El programa recorre las 12 imágenes de resumen diario de la fecha indicada y extrae los valores del pixel en el cual "caen" las coordenadas geográficas definidas [21].

[2] https://github.com/INTA-Radar

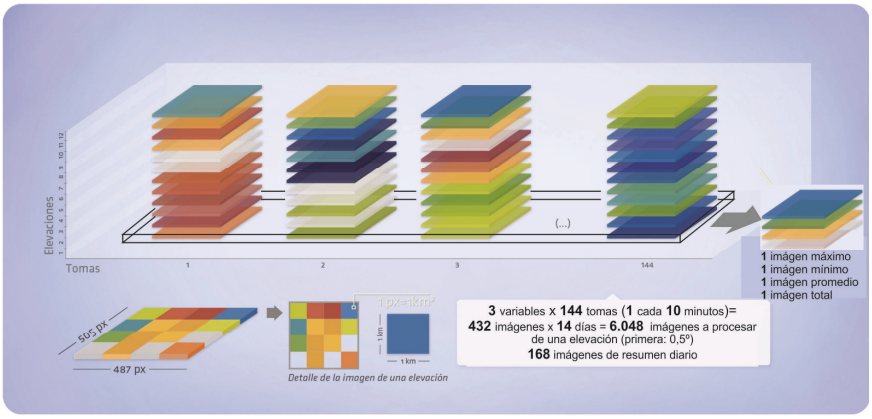

Fig. 1. Esquema del cálculo de variables de radar de resumen diario.

Tabla 1. Variables derivadas de los datos polarimétricos (Z, Z_{DR} y Rho_{HV}) del radar.

Variable	Descripción
MxDbz1	Máximo de Z ocurrido durante las 24 horas en la 1ra elevación.
MnDbz1	Mínimo de Z ocurrido durante las 24 horas en la 1ra elevación.
AvDbz1	Promedio de Z ocurrido durante las 24 horas en la 1ra elevación.
TotDbz1	Total de Z ocurrido durante las 24 horas en la 1ra elevación.
MxZDR1	Máximo de Z_{DR} ocurrido durante las 24 horas en la 1ra elevación.
MnZDR1	Mínimo de Z_{DR} ocurrido durante las 24 horas en la 1ra elevación.
AvZDR1	Promedio de Z_{DR} ocurrido durante las 24 horas en la 1ra elevación.
TotZDR1	Total de Z_{DR} ocurrido durante las 24 horas en la 1ra elevación.
MxRho1	Máximo de Rho_{HV} ocurrido durante las 24 horas en la 1ra elevación.
MnRho1	Mínimo de Rho_{HV} ocurrido durante las 24 horas en la 1ra elevación.
AvRho1	Promedio de Rho_{HV} ocurrido durante las 24 horas en la 1ra elevación.
TotRho1	Total de Rho_{HV} ocurrido durante las 24 horas en la 1ra elevación.

Se analizaron las estadísticas básicas del *dataset* y se realizaron diagramas de caja comparando los casos negativos y positivos usando R. Se aprecian diferencias entre las clases, siendo más evidente en algunas variables como los valores máximos y promedios de Z, los valores mínimos y máximos de Z_{DR} y los valores mínimos de Rho_{HV}.

3 Resultados

Se seleccionaron las técnicas RL, DT (C4.5), NB y SVM porque presentan buenos resultados en clasificar granizo en los antecedentes. Se optó por C4.5 para el DT porque es el más utilizado en los trabajos previos. En el caso de RL, DT y NB se presume que facilitaran el análisis de las variables seleccionadas por los modelos; se utilizó Tanagra [25] para correr estas técnicas. Se aplicó *Gene Expression Programming* (GEP) porque aparece como una buena herramienta en la clasificación de imágenes

satelitales (Ej:[26],[27]) e interesa analizar su comportamiento con imágenes de radar; se evolucionó una RL porque es la técnica más utilizada en los antecedentes. Se usó GeneXproTools 5.0 para su ejecución. La tabla 2 resume los parámetros utilizados en cada algoritmo.

Tabla 2. Parámetros de configuración de las técnicas utilizadas.

Técnica	Parámetros
GEP	Función objetivo: RL, Función Fitness: Máxima Verosimilitud, Cromosomas: 30, Genes: 4, Tamaño de Gen: 32, Linking Function: Addition, Estrategia: Optimal Evolution, Conjunto de funciones: 29, Operadores Genéticos: 32. Estos parámetros son los sugeridos por [28] de acuerdo a la cantidad de variables de entrada.
RL	Función de costo: Máxima Verosimilitud. No se estandarizaron ni normalizaron las variables continuas. Corte: 0,5.
DT (C4.5)	Max. .Nro de Hojas: 5, Nivel de confianza: 0,25, Criterio de división: Gain Ratio. Sin prunning.
SVM	Exponente:1, Filtro: normalizado.
NB	Lambda: 0.0, se asume homocedasticidad.

El problema se trató de forma binaria con todas las técnicas. El dataset se dividió aleatoriamente en una proporción utilizada por otros autores ([3],[4],[11],[12]) que consiste en usar 2/3 de los casos para entrenamiento y 1/3 de los casos para validación, manteniendo la proporción original de casos positivos y negativos en cada set de datos.

Para medir la performance de cada modelo generado se calcularon las medidas Probability Of Detection (POD) o *Recall*, False Alarm Ratio (FAR) y Percent Correct (PC) o *Accuracy;* sobre el dataset de validación, ya que son las medidas más utilizadas en los antecedentes. Las formulas 1 a 3 presentan el cálculo de cada medida. La tabla 3 presenta los resultados de cada técnica y la Fig.2 muestra detalles de los modelos obtenidos.

$$\text{POD} = \text{Verdaderos Positivos} / (\text{Verdaderos Positivos} + \text{Falsos Negativos}) \quad (1)$$

$$\text{FAR} = \text{Falsos Positivos} / (\text{Verdaderos Positivos} + \text{Falsos Positivos}) \quad (2)$$

$$\text{PC} = (\text{Verdaderos Positivos} + \text{Verdaderos Negativos}) / (\text{Verdaderos Positivos} + \text{Verdaderos Negativos} + \text{Falsos Positivos} + \text{Falsos Negativos}) \quad (3)$$

Tabla 3. Medidas de performance en validación de cada técnica utilizada.

	GEP	RL	DT (C4.5)	SVM	NB
POD	0,6942	**0,8171**	0,6857	0,8052	0,5285
FAR	0,2364	0,4463	**0,2066**	0,4876	0,4628
PC	**0,8668**	0,8541	0,8541	0,8436	0,7590

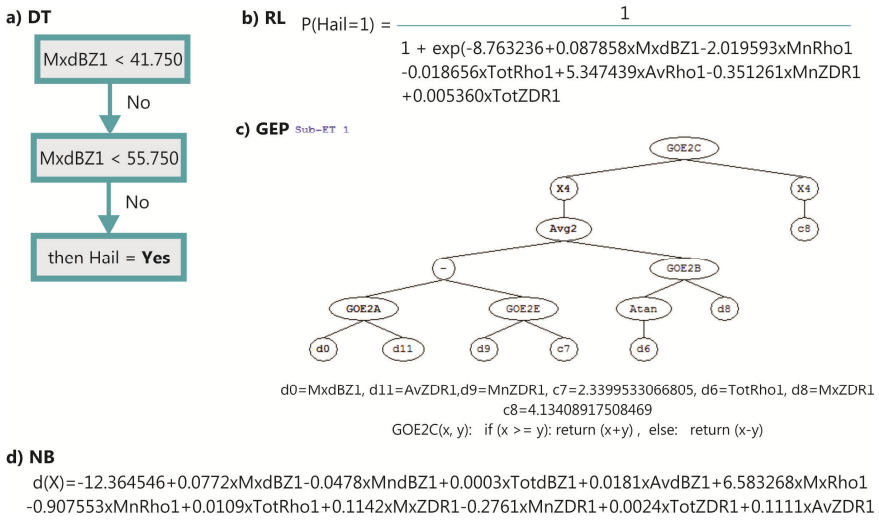

Fig. 2. a) Recorte de una de las reglas obtenidas con DT. b) Función logística para la clasificación granizo=1. c) Uno de los cuatro árboles de expresión obtenidos con GEP, ejemplo de función, valores de variables y constantes. D) Función lineal obtenida con NB.

La técnica que mejor clasifica los casos en forma global es GEP, aunque RL, DT y SVM presentan resultados similares, NB es la técnica que muestra mayor confusión. En la clasificación de granizo es necesario contar con un buen rendimiento al momento de detectar los casos positivos, el mejor valor de POD lo consigue RL, seguido por SVM. En un segundo escalón aparecen GEP y DT, siendo NB la de peor performance. Finalmente, es importante tener un valor bajo de FAR, en este aspecto DT tiene el mejor resultado, seguida por GEP. Las tres técnicas restantes presentan valores elevados de falsas alarmas.

Los coeficientes son consistentes entre los modelos NB y RL en cuanto al signo; los cuales son coherentes con los antecedentes: a mayor Z (MxDbz1), valores menores (MnZDR1) o extremos (MxZDR1, TotZDR1) de Z_{DR} y menores valores de Rho_{HV} (MnRho1, TotRho1) mayor probabilidad de ocurrencia de Granizo. La relación de valores altos de Z también se aprecia en el DT obtenido.

4 Conclusiones y Trabajos Futuros

Las técnicas que mejor relación presentan en los resultados de performance de las medidas POD, FAR y PC son GEP y DT, mientras que la de menores valores es NB.

RL, DT y NB permiten analizar de manera más sencilla, el comportamiento de las variables involucradas ante la presencia de granizo. Esta propiedad es importante para la caracterización del comportamiento de las variables del radar en la identifica-

ción de granizo en la región. Por ejemplo, a partir del DT generado, se ve que 55 dBZ es un valor más adecuado como indicador de presencia de granizo para La Pampa, en lugar de los 45 dBZ configurados por defecto en el radar, a los 50 dBZ utilizados para Paraná por [19] y los 60 dBZ indicados para Pergamino por [19].

Las técnicas supervisadas de MD resultaron ser herramientas adecuadas para generar modelos de clasificación de granizo en superficie utilizando datos polarimétricos de un radar meteorológico de banda C.

Como trabajos futuros sería importante realizar nuevas pruebas incorporando otras variables polarimétricas al dataset como Phi_{DP} y K_{DP} e integrar los datos en las 12 elevaciones y no solo de la primera. También sería interesante generar un set de datos solo con variables derivadas de Z, para evaluar el rendimiento potencial de un modelo que solo necesite esta variable ya que el mismo se podría aplicar a radares de banda C, de doble o de simple polarización (como el ubicado en INTA Pergamino).

Finalmente, sería significativo evaluar técnicas de MD para generar modelos que determinen el daño que hace el granizo en cultivos a partir de las variables polarimétricas del radar meteorológico.

Referencias

[1] E. Ponce de Leon, «Granizo». Servicio Meterologico Nacional, 1985.

[2] R. Hohl, H.-H. Schiesser, y I. Knepper, «The use of weather radars to estimate hail damage to automobiles: an exploratory study in Switzerland», *Atmospheric Res.*, vol. 61, n.º 3, pp. 215–238, 2002.

[3] L. López y J. L. Sánchez, «Discriminant methods for radar detection of hail», *Atmospheric Res.*, vol. 93, n.º 1, pp. 358–368, 2009.

[4] C. Bustos y H. Videla, «Modelo estadístico de predicción de tormentas a corto plazo para la provincia de Mendoza», en *Anales del XI Congreso Argentino de Meteorología. Catuogno, GA*, 1982.

[5] J. L. Sánchez, L. López, E. García-Ortega, y B. Gil, «Nowcasting of kinetic energy of hail precipitation using radar», *Atmospheric Res.*, vol. 123, pp. 48–60, 2013.

[6] L. López, E. García-Ortega, y J. L. Sánchez, «A short-term forecast model for hail», *Atmospheric Res.*, vol. 83, n.º 2-4, pp. 176-184, feb. 2007.

[7] M. Alexiuk, P. C. Li, N. Pizzi, y W. Pedrycz, «Classification of Hail and Tornado Storm Cells Using Neural Networks», en *1999 IEEE Western Canada Conference and Exhibition*, pp. 15–21.

[8] P. C. Li, N. Pizzi, W. Pedrycz, D. Westmore, y R. Vivanco, «Severe storm cell classification using derived products optimized by genetic algorithms», en *Electrical and Computer Engineering, 2000 Canadian Conference on*, 2000, vol. 1, pp. 445–448.

[9] L. Ramirez, W. Pedrycz, y N. Pizzi, «Severe storm cell classification using support vector machines and radial basis function approaches», en *Electrical and Computer Engineering, 2001. Canadian Conference on*, 2001, vol. 1, pp. 87–91.

[10] D. J. Gagne, A. McGovern, y J. Brotzge, «Classification of convective areas using decision trees», *J. Atmospheric Ocean. Technol.*, vol. 26, n.º 7, pp. 1341–1353, 2009.

[11] E. G. Tsagalidis, K. G. Tsitouridis, G. Evangelidis, y D. A. Dervos, «Hail Size Estimation and Prediction using Data Mining Techniques».

[12] J. Billet, M. DeLisi, B. G. Smith, y C. Gates, «Use of Regression Techniques to Predict Hail Size and the Probability of Large Hail», *Weather Forecast.*, vol. 12, n.º 1, pp. 154-164, mar. 1997.

[13] E. Collino, P. Bonelli, y L. Gilli, «ST-AR (STorm-ARchive): A project developed to assess the ground effects of severe convective storms in the Po Valley», *Atmospheric Res.*, vol. 93, n.º 1-3, pp. 483-489, jul. 2009.

[14] I. Holleman, *Hail detection using single-polarization radar*. Ministerie van Verkeer en Waterstaat, Koninklijk Nederlands Meteorologisch Instituut, 2001.

[15] C. Marzban y A. Witt, «A Bayesian neural network for severe-hail size prediction», *Weather Forecast.*, vol. 16, n.º 5, pp. 600–610, 2001.

[16] R. N. Mezher, M. Doyle, y V. Barros, «Climatology of hail in Argentina», *Atmospheric Res.*, vol. 114-115, pp. 70-82, oct. 2012.

[17] R. N. Mezher, S. Banchero, y Y. N. Bellini Saibene, «Identificación de granizo con la utilización de variables polarimétricas de los radares de Paraná y Anguil, el radar de Pergamino y daño en cultivos.», en *Congreso Argentino de Meteorología. 11. 2012 05-06 28-01, 28 de mayo al 1 de junio de 2012. Mendoza. AR.*, 2012.

[18] R. N. Mezher, L. Vidal, y P. Salio, «Hailstorms Analysis using Polarimetric Weather Radars and Microwave Sensors in Argentina», *6th Eur. Conf. Sev. Storms ECSS 2011*, 26082011.

[19] R. N. Mezher y P. A. Mercuri, «Uso de la red de radares de INTA para la detección de granizo», *XIII Reunión Argent. VI Latinoam. Agrometeorol.*, oct. 2010.

[20] J.-P. Tuovinen, A.-J. Punkka, J. Rauhala, H. Hohti, y D. M. Schultz, «Climatology of Severe Hail in Finland: 1930–2006», *Mon. Weather Rev.*, vol. 137, n.º 7, pp. 2238-2249, jul. 2009.

[21] Y. Bellini Saibene, M. Volpaccio, S. Banchero, y R. Mezher, «Desarrollo y uso de herramientas libres para la explotación de datos de los radares meteorológicos del INTA», en *XLIII Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO)-VI Congreso Argentino de AgroInformática (CAI)(Buenos Aires, 2014)*, 2014.

[22] K. Aydin, T. A. Seliga, y V. Balaji, «Remote sensing of hail with a dual linear polarization radar», *J. Clim. Appl. Meteorol.*, vol. 25, n.º 10, pp. 1475–1484, 1986.

[23] P. Bonelli, P. Marcacci, E. Bertolotti, E. Collino, y G. Stella, «Nowcasting and assessing thunderstorm risk on the Lombardy region (Italy)», *Atmospheric Res.*, vol. 100, n.º 4, pp. 503-510, jun. 2011.

[24] A. V. Ryzhkov, T. J. Schuur, D. W. Burgess, P. L. Heinselman, S. E. Giangrande, y D. S. Zrnic, «The joint polarization experiment. polarimetric Rainfall Measurement and Hidrometeor Classification», *Bull Amer Meteor Soc*, vol. 86, pp. 809–824, 2005.

[25] R. Rakotomalala, «TANAGRA: a free software for research and academic purposes», vol. 2, pp. 697-702, 2003.

[26] S. N. Omkar, N. Ramaswamy, J. Senthilnath, S. Bharath, y N. S. Anuradha, «Gene Expression Programming-Fuzzy Logic Method for Crop Type Classification», en *2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC)*, 2012, pp. 136-139.

[27] «Multi-temporal satellite image analysis using Gene Expression Programming», *Proc. Second Int. Conf. Soft Comput. Probl. Solving 2012 SocProS 2012*, 2012.

[28] C. Ferreira, «Logistic Regression Analytics Platform», *GeneXproTools Tutorials – A Gepsoft Web Resource.*, 24-oct-2013. [En línea]. Disponible en: http://www.gepsoft.com/tutorials/LogisticRegressionAnalyticsPlatform.htm. [Accedido: 01-dic-2015].