

Un Recorrido por los Principales Proveedores de Servicios de Machine Learning y Predicción en la Nube

David Corral-Plaza¹, Manuel Resinas², Juan Boubeta-Puig¹

¹Departamento de Ingeniería Informática, Universidad de Cádiz, España
{david.corral, juan.boubeta}@uca.es

²Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, España
resinas@us.es

Resumen. Los medios tecnológicos para el consumo, producción e intercambio de información no hacen más que aumentar cada día que pasa. Nos encontramos envueltos en el fenómeno Big Data, donde ser capaces de analizar esta información con el objetivo de poder inferir situaciones del futuro basándonos en datos del pasado y del presente, nos puede reportar una ventaja competitiva que nos distinga claramente de otras opciones. Dentro de las múltiples disciplinas existentes para el análisis de grandes cantidades información encontramos el Machine Learning y, a su vez, dentro de este podemos destacar la capacidad predictiva que nos proporcionan muchas de las opciones existentes actualmente en el mercado. En este trabajo realizamos un análisis de estas principales opciones de APIs predictivas en la nube, las comparamos entre sí, y finalmente llevamos a cabo una experimentación con datos reales de la Red de Vigilancia y Control de la Calidad del Aire de la Junta de Andalucía. Los resultados demuestran que estas herramientas son una opción muy interesante a considerar a la hora de tratar de predecir valores de contaminantes que pueden afectar a nuestra salud seriamente, pudiéndose llevar a cabo acciones preventivas sobre la población afectada.

Keywords: Machine Learning, Predicción, Cloud, API, Software as a Service, Big Data.

1 Introducción

En los últimos años, los usuarios se encuentran envueltos en un proceso de consumo, producción e intercambio de ingentes cantidades de información. Esta información proviene de diversas fuentes, desde redes sociales hasta dispositivos *wearables*, y gracias a las facilidades existentes hoy en día los usuarios pueden ser agentes productores o consumidores de dicha información en cualquier momento y en cualquier lugar. Vivimos en una época en la que operar con dicha información de forma eficiente nos puede proporcionar mucho conocimiento que podemos aplicar para resolver problemas específicos en cada área a tratar.

Entre las diferentes disciplinas para el tratamiento de grandes cantidades de datos, encontramos el aprendizaje automático (*Machine Learning*, ML). Aunque este término

existe desde hace mucho tiempo, en estos últimos años ha aumentado significativamente el interés de desarrolladores e investigadores a lo largo del mundo, en gran parte propiciado por el fenómeno Big Data [1] en el que actualmente nos encontramos.

El campo del ML, a su vez, engloba múltiples aplicaciones como pueden ser el reconocimiento de imágenes, detección de patrones, clasificación, detección de anomalías, predicción, etc.

En este artículo realizamos un estudio sobre aquellos proveedores de servicios de ML que ofertan la posibilidad de emplear las técnicas y algoritmos de esta tecnología con el objetivo de predecir valores. La capacidad de predecir un valor determinado o una situación que se pueda producir en función de unos datos de entrada, con un muy bajo margen de error, nos permite anticiparnos a aquello que está por suceder, resultando así una mejor competitividad dentro de cualquier ámbito de trabajo o estudio.

Landset et al. [2] realizan un estudio sobre las diferentes alternativas de ML existentes en el ecosistema de Hadoop mientras que en este trabajo nos centraremos en proveedores de servicios de ML en la nube. Al igual que el ML, el paradigma *cloud* [3] ha suscitado un gran interés tanto en empresas como en instituciones públicas y privadas. En los últimos años son muchas las grandes compañías que han pasado de ofrecer sus servicios como un producto software que debemos instalar en un equipo, a productos y servicios que se encuentran disponibles en línea, liberándonos así de restricciones ligadas a, por ejemplo, una infraestructura necesaria para poder hacer uso de dicho producto o servicio.

El resto del artículo se estructura como sigue. En la Sección 2 se presenta el concepto de ML y Predicción en plataformas de servicios en la nube. En la Sección 3, se recogen las opciones existentes hoy en día y la metodología de selección. En la Sección 4, se realiza un análisis de las opciones escogidas anteriormente. En la Sección 5, se presenta una comparativa entre las opciones previamente analizadas. En la Sección 6, se presentan los resultados obtenidos tras llevar a cabo una experimentación y su discusión. Finalmente, en la Sección 7 se presentan las conclusiones y el trabajo futuro.

2 *Machine Learning* y Predicción en la nube

Como hemos comentado anteriormente, el objetivo de este trabajo es realizar un recorrido por los diferentes proveedores de servicios de ML en la nube. Para ello, debemos tener claro qué es el ML y cuáles son los pasos o la metodología habitual a la hora de realizar un experimento de ML predictivo.

Podemos definir el concepto de ML como la aplicación de un algoritmo por parte de una máquina u ordenador para poder aprender a realizar una tarea. Esta tarea o problema a resolver puede ser de múltiples tipos, desde el reconocimiento de imágenes hasta la predicción de valores; en función de la tarea a resolver y el proveedor a emplear, el o los algoritmos que se emplean internamente variarán. En nuestro caso nos vamos a centrar en el uso de interfaces de programación de aplicaciones (*Application Programming Interface*, API) predictivas de ML, esto es la capacidad que tiene un modelo generado, a partir de unos datos de entrada, de predecir un valor concreto en función de información que le suministremos. Más específicamente, usaremos los proveedores de

servicios de ML en la nube. Entre las múltiples ventajas de esta nueva forma de trabajar en la nube con respecto a la tradicional destacamos la posibilidad de despreocuparnos de aspectos como las infraestructuras necesarias para poder ejecutar los costosos algoritmos que se emplean en los experimentos de ML. De esta forma podemos centrarnos en lo que realmente nos interesa (el estudio a realizar) y dejar que la responsabilidad de otros aspectos recaiga sobre el proveedor de servicios en la nube seleccionado. Por otro lado, cabe destacar que los usuarios con conocimientos muy básicos de ML podrán comenzar a emplear estos servicios, gracias a la usabilidad de estas plataformas, frente a la programación que se requeriría en un experimento de ML si no se usase este tipo de plataforma. Así pues, los pasos metodológicos a seguir para realizar este tipo de experimento de ML son los siguientes [4]:

1. **Fuente de datos:** en primer lugar, es importante obtener unos buenos datos de entrada. Cuanta más información contenga nuestro *dataset* más posibilidades de aprendizaje le estaremos facilitando al proveedor de servicios de ML para generar un buen modelo que pueda realizar las predicciones posteriormente. En ocasiones, entre este primer paso y el siguiente de generación de modelos, es necesario un pre procesamiento de los datos que vamos a emplear para generar los modelos, ya que son muchas las ocasiones en las que la información que vamos a emplear no se encuentra con una estructura óptima para la plataforma de ML que deseemos usar.
2. **Modelos:** el segundo paso, una vez tenemos el *dataset* preparado, es generar modelos a partir de estos datos de entrada. Los modelos pueden ser de diversos tipos, dependen fundamentalmente de los datos que contenga el *dataset* utilizado, pero los más comunes y que se suelen emplear en la predicción son: modelos de clasificación binaria, modelos de clasificación multiclase y modelos de regresión. Una vez que hemos generado los modelos, es un paso bastante común (pero no obligatorio) realizar una evaluación del modelo para determinar cuán bueno es y cuál será su rendimiento a la hora de las futuras predicciones que deseemos realizar.
3. **Predicciones:** finalmente, tendremos un modelo listo para ser empleado y sobre el cual podremos lanzar valores con el objetivo de obtener predicciones. Las predicciones pueden ser de diferentes tipos en función del proveedor, pero por norma general podemos identificar dos tipos de predicciones:
 - a. **Predicciones simples:** también conocidas como predicciones en tiempo real, son en las que introducimos una serie de valores y el modelo nos devolverá el valor predicho al instante.
 - b. **Predicciones batch:** en las que introducimos múltiples filas de valores de entrada y, para cada una de esas filas, el modelo se encargará de evaluar los valores y obtener el valor a predecir. El tiempo que tome el modelo en realizar estas predicciones en grupo dependerá del proveedor, el plan de uso que tengamos contratado y la ociosidad del sistema en tal instante.

3 Metodología del análisis comparativo

Son muchas, y cada vez más, las alternativas existentes hoy en día en el mercado de las APIs de ML. Del conjunto de opciones existentes en el mercado, hemos seleccionado

algunas de las opciones publicadas en el portal web Predictive Analytics Today [5] y otras identificadas en diferentes comparativas webs sobre proveedores de servicios de ML. Por lo tanto, las opciones consideradas son: Amazon Machine Learning, Anaconda, Apigee, BigML, Blue Yonder Platform, Dataiku, Datumbox, Google Cloud Machine Learning Engine, GraphLab Create, IBM Watson, Indico, KNIME, Logical Glue, Microsoft Azure Machine Learning, Monkeylearn, Openscoring, PredicSis, PredictionIO, y RapidMiner.

De entre este conjunto de opciones, hemos seleccionado cuatro para su estudio en profundidad. La selección de estas opciones no la hemos realizado aleatoriamente, sino que hemos llevado a cabo un proceso de revisión en la web de diferentes rankings, comparativas y análisis sobre estas opciones seleccionadas. Entre los portales empleados, destacamos KDnuggets, Data Science 101 y Quora, todos ellos portales web de gran prestigio en el ámbito del aprendizaje automático. Otro de los requisitos de selección ha sido la documentación existente y los medios de soporte para cada una de estas opciones seleccionadas.

La popularidad o dimensión de la empresa que mantiene el servicio ha sido también uno de los factores a tener en cuenta. Cuanta mayor dimensión tenga la empresa que se encarga del servicio que vamos a emplear, más posibilidades existen de poder construir y llevar a cabo nuestro proyecto a largo plazo. De igual forma, estas grandes compañías como Amazon o Google, suelen ofrecer una plataforma de desarrollo modular, en las que los diferentes servicios a emplear pueden comunicarse unos con otros; este concepto se denomina Plataforma como servicio (*Platform as a Service*, PaaS) [6], teniendo así la posibilidad de crear una arquitectura completa multifuncional. Resultando pues que las opciones a analizar sean: Google Cloud Machine Learning Engine, Amazon Machine Learning, Microsoft Azure Machine Learning, y BigML.

4 Principales APIs de ML

A continuación, presentamos un análisis detallado sobre cada una de las principales opciones previamente seleccionadas.

4.1 Google Cloud Machine Learning Engine

Cabe destacar que, a diferencia de su predecesor, no existe ocasión de poder probar estos servicios de Google de forma gratuita, ya que siempre tendremos que estar pagando en función del uso que hagamos de estos servicios.

Google Cloud Machine Learning es un servicio para desarrollar fácilmente modelos de aprendizaje automático. Se basa en el *framework* de ML desarrollado por Google: TensorFlow. Es capaz de integrarse con las diversas tecnologías de la nube de servicios de Google, de tal forma que podemos emplear este servicio junto con otros para múltiples fines como, por ejemplo, junto a Cloud Dataflow para procesar funciones, Cloud Storage para almacenar datos o Cloud Datalab para la creación de modelos.

Nos permite alta escalabilidad para generar modelos a partir de datos de diverso tamaño, de una forma distribuida y administrada con unidades de procesamiento gráfico

y unidades de procesamiento central. Dispone de una documentación muy bien estructurada, con diversas guías que nos conducirán desde los primeros pasos en la plataforma hasta la elaboración de modelos predictivos, pero en contra no suele dar muchos detalles sobre los aspectos técnicos como qué límites de tamaño de datos soporta o cuál es el abanico de modelos disponibles para generar.

4.2 Amazon Machine Learning

En Amazon Machine Learning ocurre al igual que en Google, no existe la posibilidad de probar este servicio de forma gratuita. Los costes de almacenar los ficheros generados y empleados en Amazon Machine Learning estarán asociados a los precios del servicio Amazon S3.

Con respecto a los datos, Amazon Machine Learning nos permite generar *datasets* utilizando el formato CSV. A la hora de crear modelos, en función de los datos que estemos empleado y el dato que deseemos predecir, Amazon distinguirá entre diversos tipos de modelos: clasificación binaria, clasificación multiclase, o de regresión. Igualmente, en el paso de creación de modelos, podemos personalizar diversos parámetros que pueden ser determinantes para la buena creación de un modelo. Por otro lado, para la creación de este, Amazon dividirá el *dataset* en dos porciones, una del 70% que se empleará para crear el modelo y una del 30% para entrenar y testear dicho modelo, resultando así un informe de evaluación que determina cuán bueno es nuestro modelo con respecto a estos datos de entrenamiento que le hemos proporcionado.

Finalmente, una vez tenemos el modelo listo, es el momento de empezar a realizar predicciones sobre nuestro modelo. Existen dos tipos de predicciones:

- Predicciones simples o en tiempo real: predicciones que hagamos desde el *dashboard* de Amazon Machine Learning o bien encapsulando nuestro modelo como un servicio web que pueda recibir peticiones (esto conlleva un gasto adicional).
- Predicciones *batch*: predicciones en grupos donde subimos un archivo con tantas filas como predicciones a realizar y la plataforma se encargará de ello.

4.3 Microsoft Azure Machine Learning

A diferencia de las opciones de Google y Amazon, Microsoft nos brinda la posibilidad de emplear sus servicios de forma gratuita. La creación de experimentos de ML es un tanto diferente en Microsoft Azure Machine Learning; ahora disponemos de una paleta y un área de trabajo en el que iremos arrastrando los diferentes módulos que vayamos a emplear de dicha paleta. Existen multitud de módulos con diversas funciones, desde módulos *inputs* en los que leemos o introducimos información de determinadas fuentes hasta módulos de generación de determinados modelos de aprendizaje automático.

Con respecto a las fuentes de datos, podemos usar cualquiera de los módulos disponibles para tal finalidad en la paleta (leer un fichero CSV, importar datos de un servicio web, etc.). A continuación, para la generación de los modelos, de nuevo podemos emplear algún módulo para tal fin. Cabe destacar que, a diferencia de otras plataformas, en Microsoft Azure Machine Learning podemos generar varios modelos dentro de un

mismo experimento para así evaluarlos simultáneamente y determinar cuál es el que mejor nos conviene según nuestras necesidades. Dentro de los diversos tipos de modelos existentes en esta plataforma, encontramos: modelos de detección de anomalías, de clasificación binaria y multiclase, de clusterización y de regresión. Además, para cada uno de estos tipos de modelos existen diversas alternativas o estrategias interesantes a la hora de determinar cuál es la más adecuada en nuestra experimentación.

Al contrario del resto de plataformas analizadas, esta no contempla las predicciones de la misma forma que si lo hacen Amazon Machine Learning o Google Cloud Machine Learning Engine, entre otros. Por lo cual, debemos encapsular el experimento que hemos desarrollado en un servicio web de tal forma que podemos pasarle determinados datos de entrada que hemos definido, procesar estos datos a través de este experimento diseñado, y finalmente obtener una respuesta o resultado que, en este caso, será una predicción para un valor o conjunto de valores determinado.

4.4 BigML

Finalmente, para BigML existen hasta 7 planes de pago distintos y uno gratuito. Además del soporte incluido en los planes de pago, existe la posibilidad de adquirir un soporte especial mensual.

En esta ocasión, nuestros experimentos de ML estarán limitados por el número de tareas que deseemos realizar en paralelo, así como por el tamaño máximo de los ficheros de datos a utilizar. En función del plan escogido, los límites previamente mencionados se verán extendidos. Para la opción gratuita podemos, como mucho, tener dos tareas en ejecución al mismo tiempo, mientras que en los planes de pago las tareas que podemos tener en ejecución van desde 2 hasta 64. Cabe destacar que existe la posibilidad de aplicar un 50% de descuento sobre el precio de cada uno de estos planes de pago para estudiantes, investigadores u organizaciones no gubernamentales.

Con respecto a los ficheros, en BigML podemos subir casi cualquier tipo de dato (.CSV, .TXT, .JSON, etc.) e incluso permite la opción de importar datos desde otros proveedores como Amazon S3, Dropbox o Google Cloud Storage. El tamaño máximo de estos *datasets* estará limitado en función del plan que tengamos contratado.

En cuanto a los tipos de modelos existentes, BigML contempla los que hemos estado viendo en otras opciones hasta el momento: modelos de clasificación binaria y multiclase, modelos de regresión, modelos de detección de anomalías y otros tipos de modelos como son los modelos de asociación que nos servirán para visualizar reglas de asociación entre los datos, los modelos de temas que nos servirán para desvelar temas que se den en los textos analizados o modelos de clusterización para segmentar los datos en diferentes grupos en función de su similitud.

Finalmente, en lo relativo a las predicciones, BigML ofrece tres alternativas:

- Predicciones por pregunta: insertamos valores para dos atributos del modelo seleccionado por BigML y se realizan predicciones interactivas sobre esos valores.
- Predicciones simples: en esta ocasión se insertan valores para los atributos del modelo que deseemos y se realizan predicciones interactivas sobre esos datos de entrada. Además, podemos seleccionar la estrategia de descarte.

- Predicciones *batch*: escogemos el modelo y el *dataset* sobre el cual queremos realizar el conjunto de predicciones.

5 Comparativa de APIs de ML

En esta sección presentamos una comparativa sobre los aspectos en los que hemos basado la descripción de estas plataformas: precio, *datasets*, modelos y predicciones.

Con respecto a los planes de pago de las opciones existentes, en la Tabla 1 recogemos los precios para cada una de las plataformas. Para el caso de las opciones de Amazon y Google el precio está ligado al uso de la plataforma, al igual que en Microsoft Azure pero que además le añade una cuota mensual por proyecto. En BigML existen diversos planes de pago, desde 30\$ hasta 10,000\$ mensuales.

En cuanto a la posibilidad de emplear los planes gratuitos, en la Tabla 2 contemplamos las características de BigML y Microsoft Azure Machine Learning, las dos plataformas que permiten llevar a cabo pruebas gratis.

Tabla 1. Comparativa de planes de pago

Plataforma	Plan de Suscripción
Google Cloud Machine Learning Engine	0.54\$/hora
Amazon Machine Learning	0.42\$/hora
Microsoft Azure Machine Learning	9.99\$/proyecto y 1\$/hora
BigML	30\$/150\$/300\$/2,500\$/5,000\$/7,500\$/10,000\$

Tabla 2. Comparativa de plan gratuito

Plataforma	Plan de Suscripción
Microsoft Azure Machine Learning	<i>Datasets</i> de 10GB, 1,000 transacciones API/mes
BigML	<i>Datasets</i> de 16MB, Predicciones ilimitadas hasta los 16MB, Máximo de 2 tareas en paralelo

En las Tablas 3, 4 y 5, recogemos las comparativas de estas plataformas en aspectos relacionados con *Datasets*, Modelos y Predicciones. Podemos determinar que tanto Amazon ML como BigML son dos opciones muy potentes en casi todos los campos, destacando Amazon ML en la creación de modelos mientras que BigML lo hace en el apartado de fuentes de datos. Por otro lado, Google Cloud Machine Learning Engine presenta un comportamiento bastante estable en todos los aspectos, no destacando en ninguno en particular, pero ofreciendo un buen rendimiento en todos. Finalmente, Microsoft Azure ML nos ofrece una buena experiencia en casi todos los aspectos excepto en las predicciones, ya que esta plataforma no pone en valor este aspecto, pudiéndose identificar como su punto más débil.

Tabla 3. Comparativa de fuentes de datos

Característica	Tamaño	Tipos admitidos	Personalización
Google Cloud Machine Learning Engine	---	.csv o TFRecords.	No
Amazon Machine Learning	100GB	.csv	Sí
Microsoft Azure Machine Learning	10GB	.csv, .txt, .tsv, Excel, Tabla Azure/Hive/SQL, .svmlight, .arff, .zip, .RData	Sí
BigML	16MB	.csv, .arff, .gz, .bz2	Sí

Tabla 4. Comparativa de modelos

Característica	Tamaño	Tipos	Personalización
Google Cloud Machine Learning Engine	---	Clasificación, Regresión.	Sí
Amazon Machine Learning	2GB	Clasificación, Regresión.	Sí
Microsoft Azure Machine Learning	10GB	Detección de anomalías, Clusterización, Clasificación, Regresión.	Sí
BigML	16MB	Detección de anomalías, Asociación, Clusterización, Clasificación, Regresión.	Sí

Tabla 5. Comparativa de predicciones

Característica	Tipos	Límites	Coste
Google Cloud Machine Learning Engine	Tiempo real	---	0.11\$ cada 1,000 predicciones + 0.44\$/nodo/hora
Amazon Machine Learning	<i>Batch</i> , Tiempo real	1 millón	0.10\$/1000 predicciones <i>batch</i> , 0.0001\$/predicción en tiempo real
Microsoft Azure Machine Learning	Tiempo real	---	0.5\$/1000 transacciones de API + 2\$/hora de proceso de API
BigML	<i>Batch</i> , Tiempo real, Por pregunta	16MB	En función del plan aumenta el límite de MB

6 Resultados y Discusión

Para la evaluación y prueba de experimentos de ML nos hemos decantado por dos plataformas de las previamente analizadas: Amazon Machine Learning y BigML. Para dicha experimentación hemos empleado los datos de la Red de Vigilancia y Control de la Calidad del Aire de Andalucía de la Junta de Andalucía. Debido a que los datos

no presentan una estructura óptima para su empleo en las plataformas seleccionadas, hemos pre procesado la información recibida en formato CSV a través de una clase Java. De las 61 estaciones disponibles en dicha red, hemos seleccionado la estación número 52, denominada Ranilla que se encuentra en la localidad de Sevilla y cuyas coordenadas son latitud: 37.3842476, longitud: -5.959606. Dicha estación contiene los sensores para los contaminantes Monóxido de Carbono (CO), Dióxido de Nitrógeno (NO₂), Dióxido de Azufre (SO₂), Óxido de Nitrógeno (NO), Óxido de Nitrógeno genérico (NOX), Benceno (BCN) y Tolueno (TOL). En los experimentos llevados a cabo, el NO será el campo objetivo o valor a predecir. Así pues, nuestro *dataset* está compuesto por 7 columnas, una por cada sensor, y un total de 20,890 filas, una por cada lectura válida realizada por esos sensores entre enero y mayo de 2017.

Hemos ejecutado hasta en 10 ocasiones diferentes este experimento con el objetivo de obtener unos resultados consistentes, consiguiendo las puntuaciones de los modelos de regresión entrenados por cada una de estas plataformas. En el caso de Amazon ML, se usa un algoritmo de regresión lineal, mientras que BigML emplea árboles de decisión para entrenar al modelo. Hemos llevado a cabo 999 predicciones *batch* en ambas plataformas y hemos realizado una diferencia media entre el valor predicho por la plataforma y el valor real del contaminante. Estos resultados se recogen en la Tabla 6, mientras que los costes de ejecutar cada una de estas pruebas se muestran en la Tabla 7.

Tabla 6. Resultados de la experimentación ML

Proveedor	Predicciones	Diferencia Media	RMSE Modelo
Amazon ML	999 predicciones <i>batch</i>	0.2234	8.1195
BigML	999 predicciones <i>batch</i>	0.0717	1.00

Tabla 7. Costes de la experimentación ML

Proveedor	<i>Dataset</i>	Modelos	Predicciones	Coste Total
Amazon ML	0,42\$	0,42\$	0,10\$	0,94\$
BigML	0\$	0\$	0\$	0\$

Con respecto a los tiempos necesarios para llevar a cabo esta predicción *batch*, la plataforma Amazon ML distingue dos medidas: *completion time* (2 minutos) y *compute time – approximate* (1 minuto).

Por su parte, BigML, realiza esta operación prácticamente en tiempo real, es decir, en cuanto realizamos la petición a través del *dashboard*, BigML se pone a trabajar en ella, devolviendo el resultado al instante. Desafortunadamente no existe ninguna forma de ver cuánto tiempo ha tardado BigML en realizar dicha predicción *batch*, pero según nuestra experiencia, podemos deducir que es prácticamente instantáneo.

Finalmente, podemos apreciar que ambas plataformas presentan unos buenos resultados para los modelos empleados, siendo los resultados que obtenemos en la plataforma BigML ligeramente superiores que los resultados obtenidos de Amazon ML. Sin embargo, si consiguiésemos mejorar el error de la raíz cuadrada de la media o *Root-Mean-Square Error* (RMSE) del modelo en Amazon ML, que actualmente es inferior

al del modelo en BigML, podríamos llegar a obtener mejores resultados en Amazon ML superando incluso a los obtenidos en BigML, proporcionando así unas predicciones más certeras.

7 Conclusiones y Trabajo Futuro

En este trabajo hemos realizado un recorrido por las principales API de ML. Se han seleccionado las 4 plataformas más destacadas actualmente en el mercado y hemos realizado un análisis en profundidad sobre las mismas. Finalmente hemos llevado a cabo una experimentación con datos reales de la Red de Vigilancia y Control de la Calidad del Aire de Andalucía de la Junta de Andalucía con el objetivo de predecir el valor del contaminante NO. Los resultados confirman que el rendimiento tanto de Amazon Machine Learning como BigML para realizar estas predicciones es notable e incluso cabe la posibilidad de mejorar estos resultados en la herramienta de Amazon.

Como trabajo futuro ampliaremos este estudio con la realización de la experimentación en otras plataformas como Google Cloud Machine Learning Engine y Microsoft Azure Machine Learning, y analizaremos el encapsulamiento de estos modelos de predicción en servicios web con el objetivo de integrarlos en otras arquitecturas.

Agradecimientos

Este trabajo ha sido financiado por el MINECO y Fondos FEDER (proyectos TIN2015-65845-C3-3-R y TIN2016-81978-REDT), y por el Programa de Fomento e Impulso de la actividad Investigadora de la Universidad de Cádiz.

Referencias

1. Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A., Ullah Khan, S.: The rise of “Big Data” on cloud computing: Review and open research issues. *Inf. Syst.* 47, 98–115 (2014).
2. Landset, S., Khoshgoftaar, T.M., Richter, A.N., Hasanin, T.: A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *J. Big Data.* 2, 24 (2015).
3. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. *J. Internet Serv. Appl.* 1, 7–18 (2010).
4. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* 26, 159–190 (2006).
5. Top 30 Predictive Analytics Software API in 2018, <https://www.predictiveanalytics.today.com/top-predictive-analytics-software-api/>, (2017).
6. What is PaaS? Platform as a Service | Microsoft Azure, <https://azure.microsoft.com/en-us/overview/what-is-paas/>.