## Desarrollo de una herramienta predictiva de calidad de agua para la gestión: modelos de Machine Learning + Shiny

El deterioro de la calidad del agua en Uruguay constituye un problema a nivel nacional. La Laguna del Sauce (Maldonado, Uruguay), segunda fuente de agua potable del país, no escapa a esta problemática. En este sentido, los gestores requieren de herramientas que les permitan anticiparse a condiciones no deseadas de calidad de agua (e.g. floraciones de cianobacterias, presencia de especies potencialmente tóxicas y de toxinas, la presencia de mal olor y sabor) para optimizar el proceso de potabilización y garantizar un adecuado abastecimiento a la población. La utilización de herramientas predictivas utilizando modelos de Machine Learning (ML) para la predicción de calidad del agua en el ámbito de la gestión, es hoy día muy incipiente a nivel nacional, e inclusive, a nivel internacional. La incorporación de estas herramientas acopladas a una interface interactiva, en el proceso de toma de decisión, presenta gran potencial dada su fácil utilización y costos prácticamente nulos. Es claro que la incorporación de estas herramientas en este ámbito, dependerá del buen desempeño de los modelos utilizados y de la disponibilidad de recursos humanos para su ejecución y mantenimiento.

En este contexto, el presente trabajo muestra los resultados de la generación de una aplicación en Shiny recientemente desarrollada para la predicción de atributos de calidad del agua en Laguna del Sauce usando modelos de ML. Dicha aplicación se encuentra instalada en el servidor de OSE y disponible para los gestores de la planta de OSE-UGD ubicada en Laguna del Sauce.

Dentro de los modelos de ML, se consideraron a los Random Forests (RF) dado que fueron en todos los casos los modelos que presentaron el mejor desempeño entre una amplia variedad de modelos de ML. Las variables de respuesta consideradas fueron: niveles de clorofila-a (proxy de biomasa del fitoplancton), niveles de biovolumen de cianobacterias, presencia de Grupos Funcionales Basados en Morfología (Kruk *et al*., 2010) (en concreto se consideraron los grupos III y VII que son los que comprometen más severamente a la calidad del agua) y presencia/ausencia de la especie de cianobacteria potencialmente tóxica *Microcystis aeruginosa*. Todas las variables son categóricas (2 a 3 clases) y en muchos casos desbalanceadas. Se aplicó la técnica SMOTE para lidiar con el problema del desbalance de datos en la variable de respuesta. Las variables predictoras fueron variables de calidad de agua, meteorológicas e hidrológicas. Para entrenar los modelos, se contó con una base de datos a paso diario en el período 2002-2019 (17 años de datos). La evaluación de los modelos se hizo calculando el error de clasificación sobre una muestra test independiente, no utilizada para nada en el proceso de entrenamiento. Los modelos se evaluaron no solo mediante su error global, sino que también por su capacidad de predecir las diferentes clases por separado. Cabe aclarar que los modelos predicen las variables de respuesta mencionadas del día siguiente a partir de información de las variables perdictoras del día de hoy. Las principales librerías utilizadas en este trabajo fueron: DMwR para lidiar con el desbalance de datos, randomForest para ajustar los modelos predictivos y ggplot2 para los gráficos.

Todos los modelos presentaron un desempeño global (porcentaje de casos correctamente clasificados) sobre la muestra de testeo independiente de entre 78 y 94%. El desempeño por clase también fue muy bueno, presentando en general valores

por encima del 80%, salvo algunas excepciones. En concreto, los niveles superiores de las variables (niveles que comprometen en mayor medida la calidad de agua) fueron muy bien clasificados.

Todo el desarrollo descrito, se incorporó en una aplicación Shiny donde los gestores pueden implementar los modelos en tiempo real obteniendo predicciones para el día siguiente. Además de las predicciones se brinda información gráfica para visualizar la evolución temporal de variables predictoras relevantes, así como la distribución y relación entre variables. También se encuentra accesible en la página de la aplicación una breve descripción de todo lo que la misma contiene, los resultados más relevantes de los modelos y los códigos para generarlos. Vale mencionar que un resultado que deriva del uso de la aplicación es la actualización permanente de una base de datos con todas las variables predictoras y de respuesta. Eso hace que en un futuro se puedan actualizar los modelos sin mayores esfuerzos además de tener una base actualizada disponible. Aún es muy reciente el uso de la herramienta en la planta de OSE-UGD, la evaluación de su incorporación en el proceso de toma de decisión, será un insumo muy importante para valorar la importancia de este tipo de desarrollos en el ámbito de la gestión.

La aplicación está disponible en la siguiente url:

https://matias-mw.shinyapps.io/Laguna_del_Sauce_2/S