Modelo predictivo basado en *machine learning* para la estimación de vulnerabilidades de riesgo de inundación y deslizamiento. Caso de estudio: instituciones educativas del Perú

John Wilson López Vega
11200174@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos, Lima Perú

Juan Carlos Torres Lázaro
jctorres@inaigem.gob.pe
Universidad Nacional Mayor de San Marcos, Lima Perú

José Herrera Quispe
jherreraqu@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos, Lima Perú

doi: https://doi.org/10.26439/ciis2021.5637

El fenómeno de El Niño es un evento natural que sucede cada año en el territorio peruano, este trae consigo problemas como las lluvias torrenciales que provocan inundaciones. En el territorio peruano muchas instituciones educativas son construidas sin formar parte de un estudio de suelos o vulnerabilidades como las inundaciones o deslizamientos, debido, quizás, al coste de este estudio ya que se tienen que respetar normas técnicas gubernamentales exigidas para la construcción de una entidad educativa. En vista de ello, en el presente trabajo los autores proponen un modelo predictivo basado en *machine learning* para la estimación de vulnerabilidades a partir de los datos de la zona de una institución pública. A través de esta herramienta se ha entrenado el modelo usando diversos algoritmos y datos de un *dataset* con más de 65 000 registros publicados por el Ministerio de Educación del Perú.

A PREDICTIVE MODEL BASED ON MACHINE LEARNING TO ESTIMATE FLOOD AND LANDSLIDE RISK VULNERABILITIES CASE STUDY: EDUCATIONAL INSTITUTIONS OF PERU

The El Niño phenomenon is a natural phenomenon that happens every year in Peruvian territory. It brings with it problems such as torrential rains that cause floods. Many educational institutions are built in the Peruvian territory without being part of a study of soils or vulnerabilities such as floods or landslides, perhaps due to the study's cost since they have to respect governmental technical standards required for the construction of an educational entity. Given this, in the present work, the authors propose a predictive model based on machine learning to estimate vulnerabilities from the data of the area of a public institution. Using Machine Learning, the model has been trained using various algorithms and data from a dataset with more than 65 thousand records published by the Ministry of Education of Peru.

Modelo predictivo basado en *machine learning* para la estimación de vulnerabilidades de riesgo de inundación y deslizamiento Caso de estudio: instituciones educativas del Perú

John Wilson López Vega, Juan Carlos Torres, José Herrera Quispe
11200174@unmsm.edu.pe, jctorres@inaigem.gob.pe, jherreraqu@unmsm.edu.pe

Resumen
El objetivo del trabajo de investigación es proponer un modelo predictivo basado en *machine learning* para estimar riesgo de inundación y deslizamiento en instituciones educativas del Perú, a partir de un *dataset* de riesgos con más de 65 000 registros publicados por el MINEDU. El modelo logra obtener un porcentaje de acierto del 92,407 % para el riesgo de deslizamiento y 9,612 % para el riesgo de inundación.

Introducción

El fenómeno de El Niño es un fenómeno natural que sucede cada año en el territorio peruano, este trae consigo problemas como las lluvias torrenciales que provocan inundaciones. En el territorio peruano muchas instituciones educativas son construidas sin formar parte de un estudio de suelos o vulnerabilidades como las inundaciones o deslizamientos, debido quizás al coste de este estudio, ya que se tienen que respetar normas técnicas gubernamentales exigidas para la construcción de una entidad educativa.

En vista de ello en el presente trabajo los autores proponen un modelo predictivo basado en *machine learning* para la estimación de vulnerabilidades a partir de los datos de la zona de una institución pública. Usando *machine learning* se ha entrenado el modelo usando diversos algoritmos y datos de un *dataset* con más de 65 000 registros publicados por el Ministerio de Educación del Perú.

Materiales y métodos

Dataset "Vulnerabilidad de los Locales Escolares ante la ocurrencia de Deslizamientos y/o Inundaciones originados por lluvias anómalas causadas por el Fenómeno El Niño" del MINEDU

Machine learning

Arboles de decisión

K vecinos más próximos

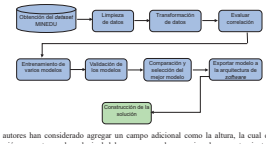
Perceptrón multicapa (MLP: MultiLayer Perceptron)

Random Forest

Microsoft Excel año del paquete Office 365

Software de *machine learning* KNIME

El proceso realizado se detalla en el siguiente gráfico.



El *dataset* utilizado posee 151 550 registros de instituciones educativas en el Perú, de los cuales muchos de ellos contenían registros nulos o vacíos, datos cualitativos y cuantitativos. Debido a la naturaleza de los campos en este trabajo se han considerado solo ciertos atributos que pueden ser tratados y categorizados y otros no, de acuerdo a la tabla 1. Posteriormente se han normalizado todos los datos a numéricos para un mejor tratamiento. Luego se ha evaluado la correlación lineal entre los campos, lo cual ha reforzado la tabla 1.

Los autores han considerado agregar un campo adicional como la altura, la cual es la elevación en metros sobre el nivel del mar, para poder apreciar el comportamiento del modelo. Dichos valores han sido alimentados con el *webservices* de Google Maps gracias a los datos de entrada que ya se tienen en el *dataset* como la latitud y la longitud. Con todos los datos ya categorizados, normalizados y filtrados por los nulos se ha obtenido un total de 65 696 registros. Los datos a pronosticar fueron riesgo de deslizamiento (0 a 5) — INGEMMET — y el riesgo de inundación ((0:NO 1:SI) — MINAM-ANA —.

Tabla 1. Variables consideradas y no consideradas del *dataset* del MINEDU

Consideradas	No consideradas
Cantidad de II. EE. que comparten el mismo local escolar, Código modalidad, Modalidad, Forma, Sexo, Código centro poblado, Área, Ubigeo, Código DRE UGEL, Lengua, d_crecen, d_juntos, d_vraem, Latitud, Longitud, Altura, Alumnos hombres, Alumnos mujeres, Total alumnos, Total docentes, Total secciones, Región hidrográfica, Unidad hidrográfica, Nombre de la unidad hidrográfica 1, Código cuenca —ANA—, Riesgo de deslizamiento (5 muy alto) —INGEMMET—, Riesgo de inundación (1 sí) — MINAM-ANA—, Movimiento de masas 1:100,000 (5 muy alta) — INGEMMET—, Inundación 1:100,000 (5 muy alta) — INGEMMET—, Infraestructura: nivel I (muy mala) —CIE—, Infraestructura: nivel II (mala) —CIE—, Infraestructura: nivel III (regular o buena) -CIE-, Categoría, Categoría final (con respuesta de UGEL), Incremento por UGEL, Categoría + UGEL + Quebrada	Código modular, Anexo, Código de local, Nombre del centro educativo, Características, Tipo de gestión, Gestión, Director, Teléfono, *E-mail*, Dirección, Localidad, Centro poblado, Región, Provincia, Distrito, DRE/UGEL, Programa, Estado, d_fte_dato, Nombre de la unidad hidrográfica 2, Nombre de la unidad hidrográfica 3, Nombre de la unidad hidrográfica 4, Nombre de la unidad hidrográfica 5, Cuencas — ANA—, Regiones de pronóstico de lluvias — SENAMHI—, Monitoreado por ODENAGED, En zona de inundación por activación de quebrada —ANA-, % de aulas para sustituir —CIE-BM

Resultados

En la tabla 4 se analizaron los modelos empleados para evaluar el mejor ajuste.

Tabla 4. ***Comparación de porcentaje de acierto***

Modelo		Riesgo de inundación
Árbol de decisión	91,42	96,808
K vecino más próximo	68,085	89,789
MultiLayer Perceptron	66,1	82,7
Random Forest	92,407	97,612

Luego de evaluar los modelos de aprendizaje por árboles de decisión, K vecino más próximo, Multilayer Perceptron y el Random Forest, se encontró que la tasa de aciertos con una división de los datos K-Folds *crosvalidation* de 10 iteraciones de los datos es del 92,407 para el riesgo de deslizamiento y de 97,612 para el riesgo de inundación, siendo estos los mejores resultados comparados con los otros modelos de *machine learning*.

Siendo el modelo de Random Forest con mejores resultados se puede realizar una proyección de las instituciones con mayores vulnerabilidades en el territorio peruano con el cual se podrían tomar mejores decisiones para elegir donde construir una institución educativa.

En la tabla 5 podemos observar que el modelo tuvo 51 033 aciertos para negativo riesgo de inundación y 13 094 aciertos para positivo de riesgo de inundación, los otros 1569 elementos contemplan los errores que tuvo el modelo.

Matriz de confución para riesgo de inundación			
		Predicción	
		NO	SÍ
Riesgo de inundación	NO	51 033	418
	SÍ	1151	13 094

Correctamente clasificados	64 127
Incorrectamente clasificados	1569
Accuracy	97,612
Error	2,388 %
Cohen's kappa	0,928

Tabla 5. Resultado de clasificación para riesgo de inundación

En la tabla 7 podemos observar que el modelo tuvo 60 708 aciertos para clasificar el riesgo de deslizamiento y 4988 desaciertos.

Matriz de confución para riesgo de deslizamiento							
		Predicción					
		1	2	3	4	5	0
Riesgo de deslizamiento	1	5840	6	21	201	0	0
	2	14	15 335	67	63	8	0
	3	72	135	20 930	1139	5	0
	4	72	80	3057	12 960	0	0
	5	0	17	6	1	5600	0
	0	24	0	0	0	0	43

Correctamente clasificados	60 708
Incorrectamente clasificados	4988
Accuracy	92,407 %
Error	7,593 %
Cohen's kappa	0,899

Tabla 7. Resultado de clasificación para riesgo de deslizamiento

Conclusiones

Se determina que el modelo generado por Random Forest es el que describe con mayor ajuste los datos analizados; este modelo presenta un porcentaje de acierto de 92,407 % para el riesgo de deslizamiento y 97,612 % para el riesgo de inundación.

El modelo no pretende reemplazar las labores del profesional encargado de llevar a cabo la decisión de construir o no la institución; solo promete ser una herramienta de apoyo para realizar más estudios. Ya que al ser una zona expuesta a vulnerabilidades se podría elegir acciones como cambiar de sede o realizar acciones que mitiguen la zona donde se construirá la institución educativa.

Referencias

- Alvarado S., Silva S. Y Cáceres D. (2010). *Modelación de episodios críticos de contaminación por material particulado (PM10) en Santiago de Chile. Comparación de la eficiencia predictiva de los modelos paramétricos y no paramétricos.* Universidad de Chile, Santiago de Chile, Chile.
- Brenes González, H. A. (2020). *Estimación de los precios de las acciones de Netflix, Inc., por medio del análisis de regresión exponencial.* Universidad Nacional Autónoma de Nicaragua Managua.
- Hidayat F. y Astsauri S. (2021). *Applied random forest for parameter sensitivity of low salinity water Injection (LSWI) implementation on carbonate reservoir. Department of Petroleum Engineering*, Universitas Islam Riau. Indonesia
- Bollado, J. Marco-Ahulló, A., Villarrasa-Sapiña, I., González, L. M. y García-Massó, X. (2018). (2018). *Dolor de espalda en estudiantes de entre 12 y 17 años: aproximación multifactorial basada en árboles de decisión*. Universidad de Valencia, Valencia, España
- Joshuva A., Sathish Kumar R., Sivakumar S., Deenadayalan, G., y Vishnuvardhan R. (2020). *An Insight on VMD for Diagnosing Wind Turbine Blade Faults Using C4.5 as Feature Selection and Discriminating through Multilayer Perceptron.*
- Kumbure M., Luukka, P., y Collan, M. (2020). *A New Fuzzy K-Nearest Neighbor Classifier Based on the Bonferroni Mean.* LUT University, Finland
- Meroni M., Waldner F., Seguini L., Kerdiles H., y Rembold F. (2021). Yield forecasting with machine learning and small data: What gains for grains? *Agricultural and Forest Meteorology. European Commission, Joint Research Centre*.

Agradecimientos

Agradezco a mi asesor José Herrera Quispe por brindarme los conocimientos necesarios que fueron pilares para poder desarrollar el presente trabajo.