PONTIFICIA UNIVERSIDAD

CATÓLICA DEL PERÚ

Facultad de Letras y Ciencias Humanas

Machine learning y cobertura de suelo en un sector del departamento de Ancash, Perú

Curso: Análisis Espacial (GEO305)

Profesores: PhD. Fabian Drenkhan y Msc. Yonatan Tarazona Coronel

Tipo de entrega: Trabajo final grupal

Alumnos:

Eylin Sánchez Callirgos (20185811)
Jose Luis Cabrera (20191061)
Naidelyn Espinoza (20180938)

2022-2

INTRODUCCIÓN

La ausencia de planificación en las ciudades de Latinoamérica es un común denominador que se ha arrastrado durante la historia de la formación de estas. Esto no solo se presenta en las ciudades sino también en los espacios rurales. Las dinámicas poblacionales así como de actividad socioeconómicas que dirigen los usos de suelo son variantes por lo que se pueden identificar en los cambios temporales de las dinámicas de uso de suelo como distribución de coberturas.

Asimismo, es importante señalar que la importancia del ordenamiento territorial para asegurar un desarrollo sostenible se ha expandido. Esto a su vez ha influido en el retorno de la puesta en valor de los instrumentos de planificación territorial. Como se explica en (Montes, P, 2001) se está revalorizando la participación de todos los actores así como la implementación de recursos dependiendo de la complejidad del territorio.

El Plan de Ordenamiento Territorial es el hilo conductor hacia un desarrollo sostenible en el que se use los recursos del espacio para generar relaciones de bienestar y conservación de toda la biodiversidad existente. Esto se aplica en todos sus niveles (incluido el grupo social que se emplace en la zona de estudio).y sus relaciones. Para ello se requiere de la identificación y conocimiento de las dinámicas territoriales que generan la diversidad de uso de suelo. Asimismo, en escenario de cambio climático se vuelve imprescindible la generación de estrategias para reducir vulnerabilidades y exposición. Para ello es vital entender las interrelaciones entre subsistemas así como sus impactos entre estos.

Para el caso del departamento de Ancash, especialmente en la zona de estudio se ha detectado la presencia de actividades económicas que generan estabilidad para las poblaciones. Asimismo, estas estructuran el espacio formando un paisaje de parches que son indicadores de la diversidad de relaciones. Específicamente las provincias que albergan el área de estudio son Huaraz, Aija, Carhuaz y Recuay. A partir de ello se pretende conocer la distribución de usos de suelo y coberturas que sirva de paisaje general en estudios integrales de planificación y ordenamiento territorial de las zonas estudiadas. Asimismo, la evaluación de resultados de una clasificación del territorio de estudio sirve para un primer acercamiento hacia la detección de problemáticas, puntos claves que generan vulnerabilidad, detección de dinámicas, efectos de actividades socioeconómicas, entre otros.

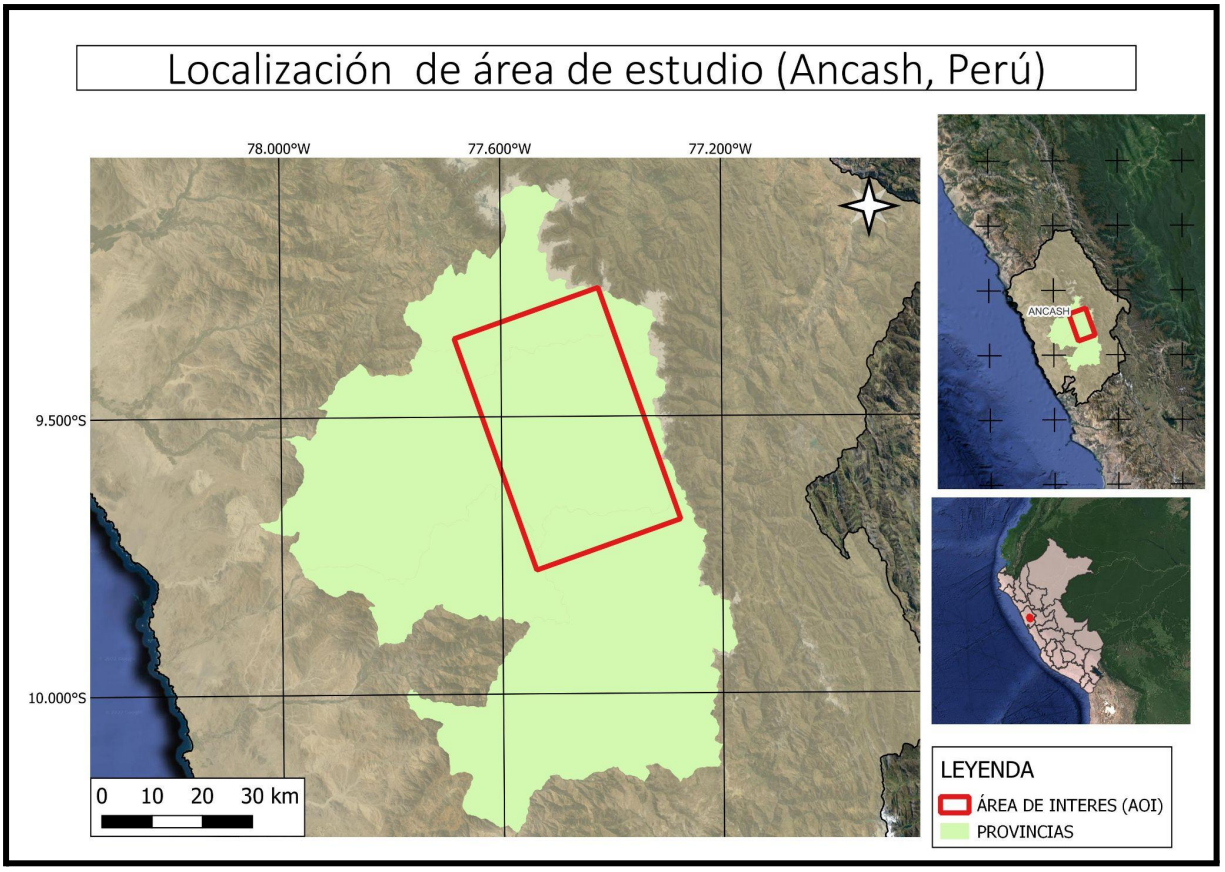


Figura 1.- Mapa de área de estudio. Elaboración propia

1. Objetivos del estudio

a. El presente trabajo pretende dar un acercamiento a los tipos de cobertura y usos de suelo de un sector de Ancash a partir del uso de herramientas de machine learning.
b. Se identificarán las distintas coberturas del área de estudio a través de una interpretación visual y posteriormente se contrastará dicha información con el uso de random forest, un algoritmo de machine learning.
c. Se clasificarán las distintas coberturas y se analizarán los resultados obtenidos, de esa manera se podrán discutir las posibles causas o efectos de los cambios y la situación actual del suelo y los recursos presentes en el área.
d. Evaluar el poder predictivo del algoritmo random forest en la clasificación de cobertura y uso de suelo mediante la construcción de una matriz de confusión, índices de precisión, así como el índice kappa.
e. Comparar el desempeño del algoritmo random forest comparando los estimadores de bondad mencionados en (d) con aquellos obtenidos mediante otros algoritmos basados en inteligencia artificial (support vector machine & neural networks)

DATOS Y MÉTODOS

Para el estudio del territorio, a través de una mirada ex situ y para una primera aproximación al conocimiento de las dinámicas principales que impactan en el mismo, una herramienta fundamental es la clasificación de usos de suelo. En primer lugar, resulta vital entender lo importante del LULCC así como su relación con la gestión territorial.

Por sus siglas en inglés, LULCC hace referencia a los cambios de uso y cobertura de suelo. Esta clasificación "consiste en categorizar las cubiertas terrestres en diferentes tipos según su uso y cobertura, como cultivos, bosques, caminos, áreas residenciales o industriales" (Wang et al., 2022). En el caso del presente estudio, siendo una zona andina del Perú, los tipos de cobertura pueden ser diversos, considerando la heterogeneidad del territorio. Asimismo, durante los últimos años, las mejoras en la tecnología de sensoramiento remoto ha mejorado y "el libre acceso a imágenes de alta calidad especialmente diseñadas para el estudio de la vegetación…ha permitido aumentar la precisión global de la clasificación de LULC (Trujillo-Jiménez et al., 2021). Sin duda, analizar los cambios en el uso de la tierra es importante para múltiples campos, especialmente es útil para los tomadores de decisiones a distinta escala y las políticas que puedan formular a partir de los estudios obtenidos de la dinámica física, aunque el LULCC constituye más que ello. Como menciona Wang y otros, la identificación de los cambios en el uso de la tierra, permite explicar las causas y efectos de los procesos de cambio en el suelo (2022). Pero ello debe ir acompañado de información que explique el contexto del área en cuestión, ya sea en términos ambientales, sociales, económicos, etc.

Ahora bien, para generar los mapas de clasificación y cambios de uso de suelo, de debe partir del uso de imágenes satelitales y datos remotos, para ello existen diferentes métodos, por ejemplo, clasificadores supervisados de aprendizaje automático, basado en clasificación de pixeles (Trujillo-Jiménez et al., 2021). En el caso del presente trabajo, se planea usar el logaritmo de machine learning llamado Random Forest o Bosques Aleatorios, que pertenece a la clasificación supervisada y se usa para asignar con exactitud a los datos de entrenamiento las categorías correspondientes de cobertura de suelo. Sin embargo, a pesar de los avances en el tema, aún hay desafíos que persisten y nuevos retos que afrontar relacionados con el creciente volumen de datos y la necesidad de capacidad computacional. A pesar de ello, en palabras de Wang y otros, se cree que machine learning tiene potencial y un futuro prometedor en cuanto a la incorporación de nuevas variables en el uso de LULCC a través del acceso a mayores cantidades de data no in situ y mejora de algoritmos transitorios (2022).

a) Datos

Se ha seleccionado una imagen satelital Sentinel 2, que cuenta con las bandas B2, B3, B4, B8, B11 y B12. Asimismo, el área de interés mostrado es la ciudad de Huaraz y alrededores en el departamento de Ancash en Perú con fecha del 26 de agosto del 2022. Algunas especificaciones de las imágenes que provee este satélite, que vuela en una órbita sincrónica al sol a 786 km de altura, es que cuentan con una resolución(res.) espacial de 10 a 60 metros, una res. temporal de 5 días, de una res. espectral de 13 y una res. radiométrica de 12 bit.

b) Algoritmos de inteligencia artificial

Todos los algoritmos empleados fueron ejecutados a partir de la librería scikeo de Python.

Random Forest

Este algoritmo ha sido uno de los usados en el presente trabajo debido a su precisión, competitividad y bajo costo computacional para la obtención de resultados (Ordoñez et al., 2020). Se debe saber que este método es “un meta-clasificador que utiliza árboles de decisión como clasificadores base, donde cada clasificador contribuye con un voto para la asignación de la clase más frecuente” (Del Toro Espín et al., 2015). Asimismo, es considerado, a partir de la vasta literatura existente sobre el tema, uno de los más conocidos y utilizados algoritmos para distintas investigaciones donde se utilizan datos de teledetección (Trujillo-Jiménez et al., 2021). Además, este método de clasificación es relativamente práctico ya que solo usa dos parámetros, “el parámetro m (número de variables utilizadas al azar para cada división o número de variables predictivas) y el parámetro k (número de árboles de clasificación)” (Del Toro Espín et al., 2015). Otra ventaja más es que su rendimiento es mejor cuando se trabaja con gran volumen de datos de entrenamiento y estos son “homogéneos y equilibrados” (Trujillo-Jiménez et al., 2021).

Support Vector Machine

El método de Support Vector Machine, de acuerdo a Oo y otros (2022, pp. 4-5), es un algoritmo no paramétrico que crea hiperplanos basados en las mayores brechas posibles encontradas en el conjunto de los datos de entrenamiento. Una vez identificados los hiperplanos, estos forman divisiones que permiten categorizar los objetos segmentados en alguna de las clases de uso y cobertura de suelos. Los autores muestran, además, mediante revisión de literatura y su propio estudio que, junto al algoritmo Random Forest, Support Vector Machine es el que brinda una mejor performance integral (menor error, menor tiempo de ejecución y mayor precisión) al clasificar uso y cobertura de suelos.

Multi-layer Perceptron Classifier

Según Wang y otros (2022), las redes neuronales artificiales son artificios matemáticos que emulan el sistema neurológico humano. Se emplean neuronas artificiales conectando los inputs y outputs mediante ciertos pesos. Cada neurona es un perceptrón que toma decisiones de clasificación de información en base a regresiones estadísticas. Estas regresiones generan señales que son empleadas por las neuronas como una función predictora no necesariamente lineal. Cada señal es combinada según los pesos correspondientes, generando así un resultado producto de la interconexión ponderada entre neuronas. Los autores reconocen que esta estructura neuronal del algoritmo cuenta con propiedades como la auto-organización, adaptación y auto-aprendizaje las cuales brindan ventajas en tareas como la clasificación de data multi-hiperespectral (2022, p. 6).

c) Variables de interés adicionales

1. NDVI

El Índice de vegetación de diferencia normalizada permite observar la densidad y la vigorosidad de la vegetación. Así se puede saber su condición actual y los cambios que ha sufrido a lo largo de los años a partir de un análisis multitemporal, por ejemplo. Por lo general, vegetación más robusta es de color verde, pero vegetación en mal estado tiende a tener tonos verdes más oscuros. Sin embargo, siempre se debe considerar el tipo de vegetación y otras consideraciones (estacionalidad, la fase de crecimiento de la planta, etc.) ya que a partir de eso pueden existir variaciones significativas.

2. NDSI

El Índice Diferencial Normalizado de Nieve permite identificar superficies glaciares y nevados. En una zona como Huaraz, especialmente hacia el este, la presencia de masa glaciar en los Andes es significativa, por ese motivo se hizo uso de este índice.

d) Procedimiento

1. Clasificación visual de clases de cobertura de suelo a partir de la imagen satelital
2. Creación de 432 puntos vectoriales en cada superficie y categorizarlas de acuerdo a la clasificación visual ya realizada.
3. Cálculo de los índices NDVI y NDSI empleando ArcMap 10.8
4. Construcción de un raster compuesto por las múltiples bandas de la imagen satelital proveída así como los índices calculados en el paso anterior
5. Subida de los archivos producto de los pasos anteriores a drive

6. Importación de la imagen y el data frame de los puntos de clasificación en google colab empleando las paqueterías *rasterio* y *dbfread*.
7. Construcción de los modelos de Machine Learning empleando la función MLA de la librería *scikeo*. El algoritmo de Random Forest se produjo empleando la función MLA.RF; el método de Support Vector Machine, mediante MLA.SVM; finalmente, el modelo de redes neuronales se obtuvo con la función MLA.NN. Para todos los modelos se dividieron los puntos de clasificación manual entre un 70% de data de entrenamiento y el restante 30% para el test.
8. Exportación de las imágenes resultantes de la clasificación supervisada y procesamiento mediante ArcMap para su adecuada presentación en mapas.
9. Presentación de los mapas y los índices pertinentes de cada modelo (exactitud y kappa), así como las matrices de confusión.

Cobertura	Etiqueta de clase	Color de clase
Urbano	1	
Cultivos	2	
Cuerpos de agua	3	
Suelo desnudo	4	
Minería	5	
Glaciar	6	
Bosque	7	

Tabla 1. Clases identificadas. Elaboración propia.

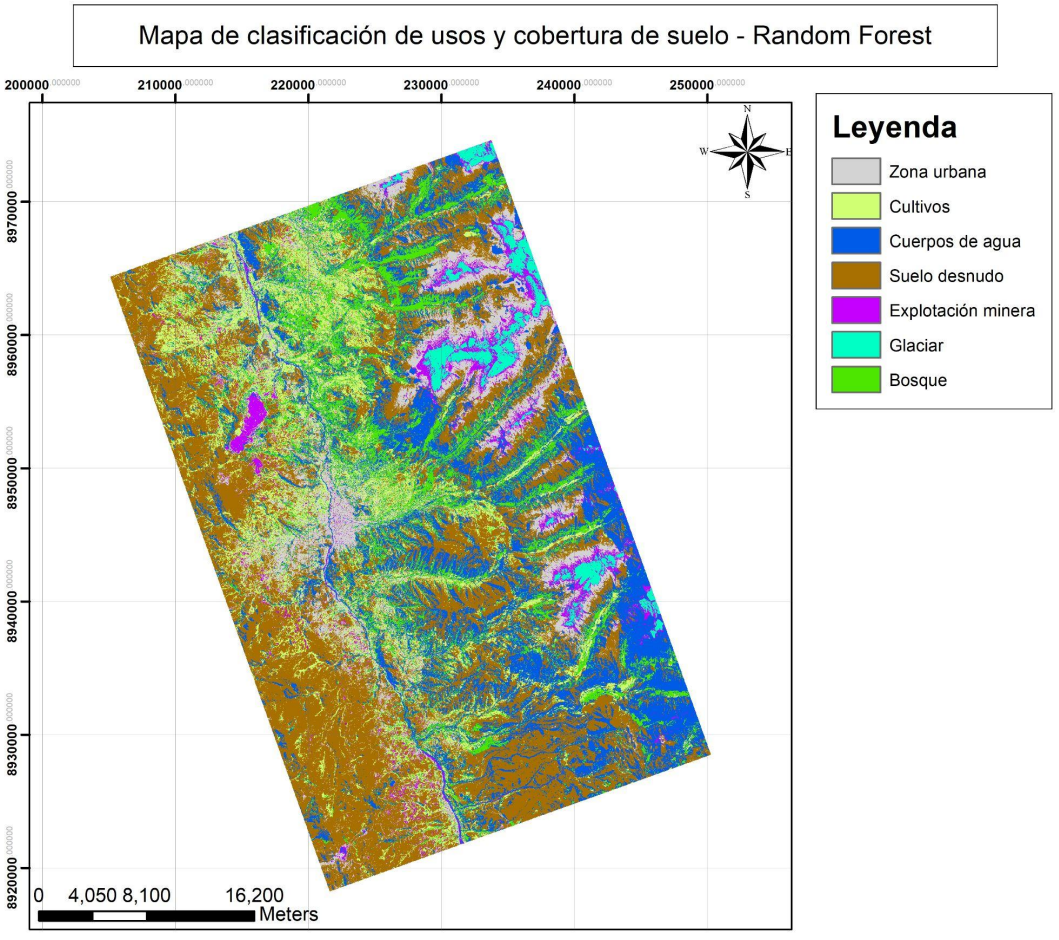
RESULTADOS Y DISCUSIÓN

Después de haber completado la información para ejecutar el modelo, se obtuvo la siguiente clasificación.

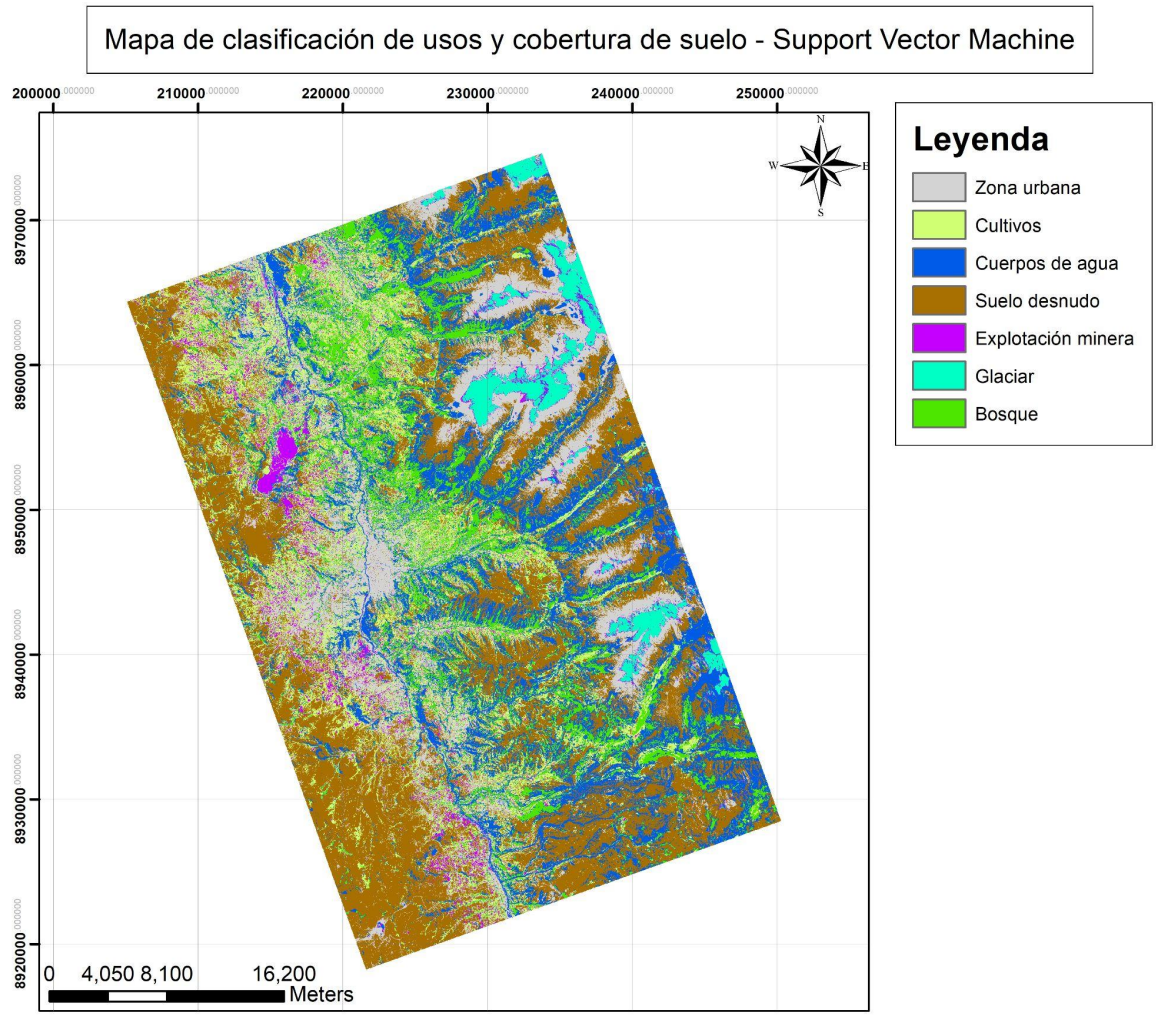
a) Random Forest

	0	1	2	3	4	5	6	Total	User's Accuracy	Commission
0	12	1	3	1	1	0	0	18	66.67	33.33
1	0	16	0	0	0	0	0	16	100	0
2	2	0	17	2	1	0	0	22	77.27	22.73
3	2	0	0	19	0	0	0	21	90.48	9.52

4	**0**	**0**	**0**	**1**	**11**	**1**	**0**	**13**	**84.62**	**15.38**
5	**0**	**0**	**0**	**0**	**1**	**23**	**0**	**24**	**95.83**	**4.17**
6	**0**	**1**	**0**	**0**	**0**	**0**	**15**	**16**	**93.75**	**6.25**
Total	**16**	**18**	**20**	**23**	**14**	**24**	**15**	**NA**	**NA**	**NA**
Producer's accuracy	**75**	**88.89**	**85**	**82.61**	**78.57**	**95.83**	**100**	**NA**	**NA**	**NA**
Omission	**25**	**11.11**	**15**	**17.39**	**21.43**	**4.17**	**0**	**NA**	**NA**	**NA**

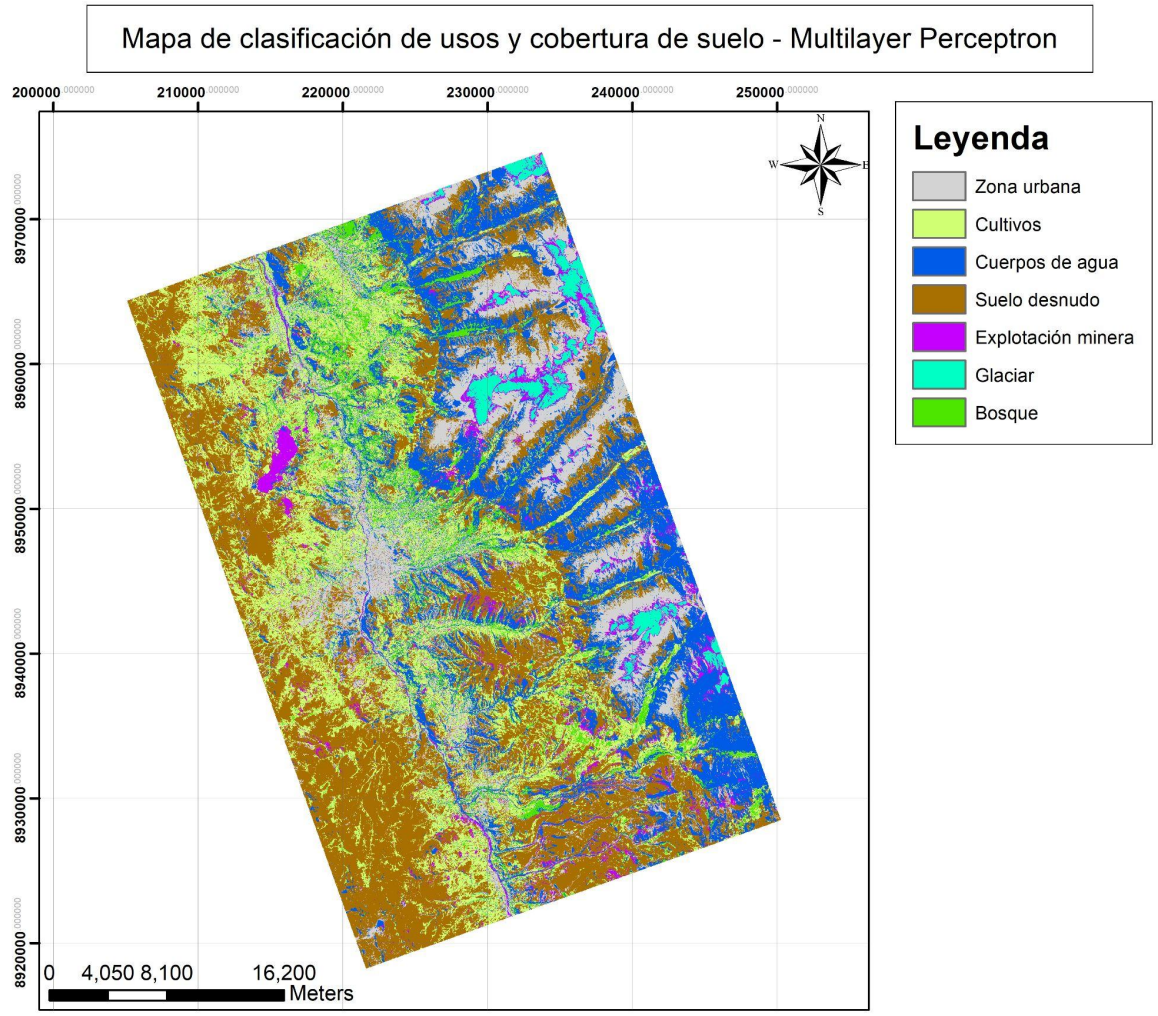


b) Support Vector Machine



	0	1	2	3	4	5	6	Total	User's Accuracy	Commission
0	13	2	0	1	5	1	0	22	59.10	40.90
1	2	16	0	1	0	0	0	19	84.21	15.79
2	0	0	15	2	1	0	1	19	78.95	21.05
3	1	0	0	16	0	0	0	17	94.12	5.88
4	1	0	1	0	14	0	0	16	87.5	12.5
5	0	0	0	0	0	20	0	20	100	0
6	0	1	0	0	0	0	17	17	100	0
Total	17	18	16	20	20	21	18	NA	NA	NA
Producer's accuracy	76.47	88.89	93.75	80	70	95.24	99.44	NA	NA	NA
Omission	23.53	11.11	6.25	20	30	4.76	5.56	NA	NA	NA

c) **Multi-layer Perceptron Classifier**



	0	1	2	3	4	5	6	Total	User's Accuracy	Commission
0	12	0	1	2	1	1	0	17	70.59	29.41
1	1	19	3	2	0	0	3	28	67.86	32.14
2	0	0	11	0	2	0	4	17	64.71	35.29
3	2	0	1	12	0	0	0	15	80	20
4	1	0	5	0	12	0	0	18	66.67	33.33
5	0	0	0	0	0	17	0	17	100	0
6	0	1	0	0	1	0	16	18	88.89	11.11
Total	16	20	21	16	16	18	23	NA	NA	NA
Producer's	75	95	52.38	75	75	94.44	69.57	NA	NA	NA

	0	1	2	3	4	5	6	Total	User's Accuracy	Commission
accuracy										
Omission	**25**	**5**	**47.62**	**25**	**25**	**5.56**	**30.43**	**NA**	**NA**	**NA**

d) Comparación de los modelos

	Support Vector Machine	Random Forest	Multilayer Perceptron
Overall accuracy	0.85385	0.86923	0.76154
Kappa	0.82958	0.84656	0.72136

Tras realizar un análisis de los resultados obtenidos con los tres tipos de clasificación de usos de suelo se identifica que el método de Random Forest presenta una mayor precisión obtenida. Esta es de un 86.92%, es decir, la precisión es aceptable para la evaluación de clases de suelo y genera un confianza media en las clases representadas y diferenciadas. En otras palabras, se podría usar la clasificación para diferenciación de usos de suelos sin la necesidad de ir a campo, claramente no es igual a la realidad pero brinda una primera aproximación confiable.

El segundo clasificador con mayor precisión general es el de SVM. Este consta de un 85.3% de precisión. Finalmente, el tercer clasificador en confiar es el de Multilayer Perceptron. En este último la precisión es de 76.15%. Por otro lado, si se compara el índice de Kappa se puede identificar que existe una mayor coincidencia entre los valores de entrenamiento y valores ya clasificados para el caso del clasificador RF. Este es de un 84,6%. En segundo lugar, el siguiente clasificador que genera mayores coincidencias para ambos parámetros es el de SV con un 82.9%. Finalmente, se encuentra el clasificador Multilayer Perceptron que genera una menor concordancia entre valores de entrenamiento y los clasificados obtenidos como resultado. Este último consta de 72.1%.

El análisis explicado se enmarca dentro de una comparación general del producto obtenido tras la clasificación. Sin embargo, dependiendo de la clase específica no necesariamente el modelo con mejor precisión general lo clasifica de la manera más parecida a la realidad. Para el caso de la diferenciación de clases desde un análisis visual se observa de manera directa que, por ejemplo, las zonas de uso de suelo empleada para explotación minera resulta más acorde con la realidad con el resultado obtenido con el clasificador Support Vector Machine que con los otros modelos.

Respecto a la clase de cuerpos de agua sucede lo mismo, el modelo SVM distingue mejor esta clase de las otras. Ello se refleja sobre todo en el flanco oriental de la imágen. De forma más precisa, en la zona cercana a las montañas que contiene la zona de glaciares. Tanto el modelo RF como el de Multilayer Perceptron generan mayor área de uso de suelo concentrado y casi adyacente de manera continua, lo cual no coincide con la realidad. Por otro lado, a nivel de la matriz de confusión, se puede observar que; en RF, las clases de cultivos y glaciares han obtenido la mejor precisión; en SVM, las clases glaciar y bosque tienen las mejores precisiones y en Multilayer Perceptron, la clase glaciares es la que cuenta con menos errores.

Asimismo, para la clase de zona urbana existe una distribución con mayor precisión para el caso del resultado obtenido con RF. En esta se observa menor error respecto a la clasificación de las zonas urbanas cercanas a las zonas glaciares. Ese es un error en menor proporción fácilmente identificable si se compara con los otros dos modelos. Los otros clasifican una mayor área de zona urbana cercana a glaciares lo cual no coincide con la realidad.

CONCLUSIONES

Se puede observar que de los tres modelos de clasificación empleados, el que presenta menor error visualmente, en primera instancia, es el de Support Vector Machine. Esto se deduce a partir de la comparación visual entre la imágen clasificada final y la imagen en composición natural en conjunto con las variables de apoyo como son el NDVI y NDSI. A nivel técnico, el modelo que presentó menos errores ha sido el realizado con el algoritmo Random Forest. Asimismo, este último tiene una mayor precisión.

Por último, se considera que el uso de los índices espectrales y su inclusión en los modelos desarrollados han sido una herramienta que ha añadido mayor información para obtener una mejor visualización en la clasificación final. En futuros trabajos se podría hacer uso de otras variables como información de textura obtenida a partir de funciones semivariograma (Del Toro Espín et al., 2015, p.332-333) o altitud obtenida a partir de un Modelo Digital de Elevación para aumentar la información y la precisión del modelo seleccionado.

REFERENCIAS

Del Toro Espín, N., Gomariz-Castillo, F., Cánovas-García, F., & Alonso-Sarría, F. (2015). Comparación de métodos de clasificación de imágenes de satélite en la cuenca del río Argos (Región de Murcia). Boletín de la Asociación de Geógrafos Españoles.

MINAM. (2016). *Instrumentos Técnico-Normativos del Ordenamiento Territorial*.

Montes, P. (2001). *El ordenamiento territorial como opción de políticas urbanas y regionales en América Latina y el Caribe medio ambiente y desarrollo*. *5*.

Oo, T. K., Arunrat, N., Sereenonchai, S., Ussawarujikulchai, A., Chareonwong, U., & Nutmagul, W. (2022). Comparing Four Machine Learning Algorithms for Land Cover Classification in Gold Mining: A Case Study of Kyaukpahto Gold Mine, Northern Myanmar. Sustainability, 14(17), 10754.

Ordóñez, C. M., Ordóñez, J. M., Fierro, L. P., & Casas, A. F. (2020). Mapeo de cobertura terrestre utilizando aprendizaje máquina. Investigación e Innovación en Ingenierías, 8(3), 85-101.

Trujillo-Jiménez, M. A., Liberoff, A. L., Pessacg, N., Pacheco, C., & Flaherty, S. (2021). Metodología de clasificación automática de uso y cobertura de suelo. In *XIII Congreso de AgroInformática (CAI 2021)-JAIIO 50 (Modalidad virtual)*. http://bvpad.indeci.gob.pe/doc/estudios_CS/Region_Ancash/ancash/huaraz.pdf

Wang, J., Bretz, M., Dewan, M. A. A., & Delavar, M. A. (2022). Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Science of The Total Environment*, 153559.