Identificación Inteligente de Cultivos Estivales mediante sensores remotos

Castillo Cristina[1][0009−0007−5162], Veramendi Brenda[1][0009−0000−5495−4246], and Revollo Sarmiento G. Noelia[1,2][0000−0002−1532−5428]

[1] Facultad de Ingeniería, Universidad Nacional de Jujuy (UNJU). Ítalo Palanca 10, 4600. San Salvador de Jujuy, Argentina.
cristinadv3582@gmail.com, brendaveramendi@gmail.com
[2] Instituto de Ecorregiones Andinas, CONICET - UNJU. Canónigo Gorriti 237, Y4600. San Salvador de Jujuy, Argentina. grevollo@fi.unju.edu.ar

Resumen El sector agrícola viene experimentando un profundo cambio hacia la transformación digital, buscando anticiparse a eventos edafoclimáticos que puedan inferir en los resultados productivos y facilitar la toma de decisiones de forma remota. El monitoreo de cultivos con sensores remotos y la aplicación de técnicas de aprendizaje de máquina son herramientas que permiten identificar cultivos, cambios en la fenología, anomalías y a menor o casi nulo costo que los métodos tradicionales. La identificación de cultivos, obtención de estimaciones y predicciones de rendimiento son esenciales para mejorar las prospectivas productivas del país. En esta primera etapa de investigación se clasificó maíz vs. soja en imágenes multiespectrales Sentinel-2 con datos registrados de la campaña 2019/2020 del departamento General López (Santa Fe). Dos algoritmos de aprendizaje de máquina: Random forest (RF) y Support vector machine (SVM) fueron implementados con una precisión global de 81 y 74 %, respectivamente. El objetivo general es extender estos resultados a una mayor región agroproductiva.

Keywords: Sensores remotos · Machine learning · Clasificación.

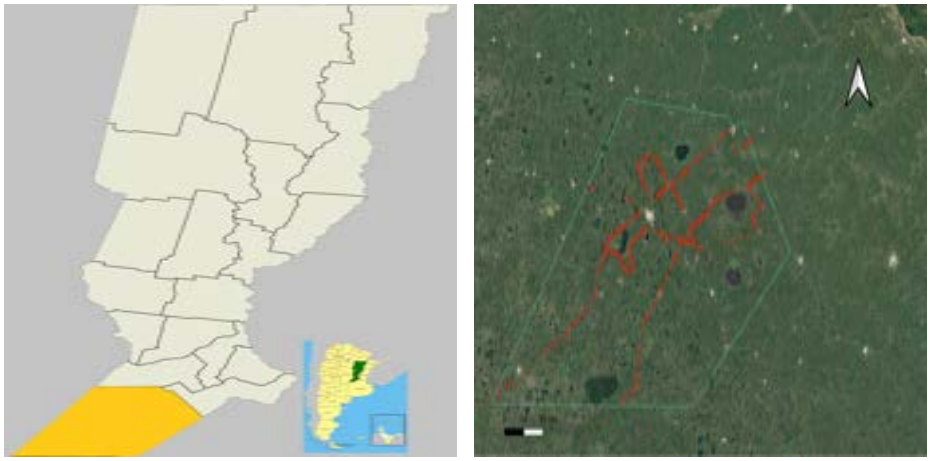
1. Introducción

La agricultura es la fuente básica de suministro de alimentos de todos los países del mundo. El aumento de la oferta por parte del sector agrícola tiene una gran importancia para el desarrollo económico de un país [1]. La información provista por sensores remotos es importante para estudios relacionados con el uso y ocupación de la superficie del suelo [2]. Los avances tecnológicos en el procesamiento digital de imágenes satelitales, junto con técnicas de aprendizaje automático, aporta nuevos métodos de análisis y seguimiento de la producción agrícola [3]. Estos métodos tienden a producir una mayor precisión en comparación con clasificadores paramétricos tradicionales, especialmente para datos complejos con múltiples variables de predicción [4]. El objetivo del presente trabajo es desarrollar e implementar un método robusto de clasificación de cultivos estivales implementando técnicas de aprendizaje automático y sensores remotos.

2. Materiales y métodos

2.1. Área de estudio y muestreo

La zona de estudio abarca una superficie de 11.558 km en del Departamento General López, Provincia de Santa Fe (Fig. 1). Es una región geomorfológica, denominada área de Modelado Eolico Pos-Pampeano y se caracteriza por un relieve con lomadas y zonas deprimidas, en parte cubiertas por cuencas lagunares. La actividad económica fundamental de esta región es la agropecuaria. Los principales cultivos que se producen son trigo, maíz, soja, girasol y pasturas. Los datos de muestreo fueron relevados por la Bolsa de Cereales de Buenos Aires y se corresponden a las campañas 2018/2019 y 2019/2020 [5]. El relevamiento de los lotes se realizó considerando los cultivos cercanos a la ruta, por esta razón es la variablidad en cantidad. La información obtenida está compuesta por 849 puntos, etiquetados con diferentes coberturas de suelo, cantidad y campaña (Tabla 1).



(a) Proviancia de Santa Fe. (b) Puntos de muestreo en General López.

Figura 1: Ubicación geográfica de la zona de muestreo.

En esta primera etapa de investigación, se consideraron los datos correspondientes a la campaña 2019/2020, teniendo en cuenta que durante la campaña 2018/2019 no se tuvieron buenos rendimientos dadas las altas precipitaciones ocurridas durante ese periodo . Las coberturas de suelo a identificar se corresponden con las clases mayoritarias (soja y maíz), más cuerpos de agua y zonas urbanas. Los cuerpos de agua y zonas urbanas se relevaron por cercanías a los cultivos y fueron digitalizados visualmente con el software de información geográfica QGIS. El dataset final está formado por 131 lotes de Maíz de 1º y 2º, 277 de Soja de 1º y 2º, 71 de cuerpos de agua y 90 de zonas Urbanas, un total de 542 muestras (Tabla 2).

Tabla 1: Tipo y cantidad de cultivos que forman el dataset inicial.

Cultivo	2018/2019	2019/2020
Soja de 1º	155	189
Soja 2º	1	88
Maiz 1º	82	128
Maiz 2º	1	3
Trigo	0	2
Girasol	1	0
Sorgo	6	0
Forrajes	10	45
Campo Natural	25	57
Barbecho	0	6
Agua	1	1
Afalfa	2	0
No Sabe	9	25
Urbano	0	12
Total	293	556

Tabla 2: Categorias y cantidad de cultivos del dataset

Clase	Categoria	Cantidad
1	Maiz (M)	131
2	Soja (S)	277
3	Agua (A)	71
4	Urbano (U)	90
Total (T)		542

2.2. Preprocesamiento de datos

Inicialmente, se realizó un control visual y se realizó una corrección geográfica a los puntos, moviendo así la ubicación de las coberturas de suelo a su correcta correspondencia geográfica y convirtiendo los datos a formato vectorial. Se utilizó el software libre QGIS (Sistema de Información Geográfica). Esta nueva capa de puntos posibilitó la identificación de las coberturas de suelo sobre las imágenes satelitales. El procesamiento de imágenes se realizó utilizando la plataforma de Google Earth Engine (GEE) [6]. Se utilizaron imágenes multiespectrales Sentinel-2, las mismas fueron filtradas por fechas desde la siembra a la cosecha, es decir del 1/09/2019 al 30/04/2020 y con una cobertura nubosa inferior al 10 %. La colección de imágenes está compuesta por las bandas "B2", "B3", "B4", "B5", "B6", "B7", "B8", "B11" (Tabla 3). Se computaron dos índices verdes: Índice de Vegetación de Diferencia Normalizada (NDVI) y el Índice de Vegetación ajustado al Suelo (SAVI) (Ec.1 y 2). Inclusive, se calcularon imágenes promedio, máximo, mínimo y desviación estándar para todas las bandas. Finalmente se generó un "*Stack*" de bandas, en el cual un solo píxel contiene un vector de n dimensiones con valores espectrales correspondientes a las bandas

consideradas. Sobre este dataset de imágenes se procedió a la superposición de la capa vectorial generada con los puntos y a la extracción de grupos de píxeles prototipos de cada clase.

Tabla 3: Características de cada una de las bandas de Sentinel 2.

Sentinel 2			
Bandas	S2A-Long. de onda [nm]	S2B-Long. de onda [nm]	Resolución[m]
B1-Aerosol	442.7	442.2	60
B2-Blue	492.4	292.1	10
B3-Green	559.8	559	10
B4-Red	664.6	664.9	10
B5-Red edge 1	704.1	703.8	20
B6-Red edge 2	740.5	739.1	20
B7-Red edge 3	782.8	779.7	20
B8-Near infrared(NIR1)	832.8	832.9	10
B8A-Near infrared(NIR2)	864.7	864	20
B9-Water vapour	945.1	943.2	60
B10-Cirrus	1373.5	1376.9	60
B11-SWIR 1	1613.7	1619.4	20
B12-SWIR 2	2202.4	2185.7	20

$$NDVI = \frac{NIR1 - RED}{NIR1 + RED} \tag{1}$$

$$SAVI = \frac{NIR1 - RED}{NIR1 + RED + L}(1 + L) \tag{2}$$

el factor L es encargado de amortiguar la presencia del suelo a través de valores comprendidos entre 0 (para zonas con gran densidad vegetal) y 1 (para zonas con escasa densidad vegetal), por defecto se consideró a L con un valor de 0,5.

2.3. Clasificación y Métricas de evaluación

Se implementaron dos algoritmos de aprendizaje supervisado: Random Forest (RF) y Support Vector Machine (SVM). Un modelo "Random Forest" está formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. En este caso la salida es la moda de las predicciones de los árboles [7].

El algoritmo SVM está basado en teoría estadística y encuentra el límite entre clases, utilizando planos de soporte generados a partir del set datos de

Tabla 4: Matriz de Confusión

		Prediccion		
		Negativo	Positivo	Total
Actual	Negativo	Tn	Fp	An = Tn + Fp
	Positivo	Fn	Tp	Ap = Fn + Tp
	Total	Pn=Tn+Fn	—Pp=Fp+Tp	T

Tabla 5: Grado de concordancia del índice Kappa

Kappa	Grado de acuerdo
< 0.00	Sin acuerdo
0.01-0.20	Insignificante
0.20-0.40	Bajo
0.40-0.70	Bueno
0.70-1.00	Óptimo

entrenamiento. El entrenamiento de los modelos se realizó con un 70 % de los datos y un 30 % para la validación. En el modelo de RF, se seteó el parámetro $número\_arboles = 100$ y en SVM:
$Proced.\_Decision = Voting, Tipo\_SVM = C\_SVC, Tipo\_Kernel = Linear$.

El desempeño de la clasificación, se evaluó mediante la matriz de confusión (Tabla 4), en donde cada columna de la matriz representa el número de predicciones de cada clase, y cada fila representa a las instancias en la clase real:

- Verdadero positivo (Tp) si la instancia es clasificada correctamente y su clase pertenece a la positiva.
- Verdadero negativo (Tn) si la instancia es correctamente clasificada con la clase de valor negativo
- Falso positivo (Fp) cuando la instancia es clasificada incorrectamente como positiva cuando en realidad es negativa
- Falso negativo (Fn) se presenta cuando la instancia fue clasificada incorrectamente como negativa cuando en realidad es positiva

Esta matriz permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de ajustarse a los datos observados [8]. De esta matriz, se calculan dos tipos de errores: de *'omisión' (O) y 'comisión' (C)*. Un error tipo (O), es cuando se excluye un valor de la categoría que está siendo evaluada, y un error tipo (C), cuando se incluye un valor incorrectamente en la categoría que está siendo evaluada. Las mediciones de la precisión (% correcto) se llaman Precisión del Usuario y del Productor, mide errores de comisión y omisión, respectivamente. Otra métrica normalmente utilizada es el coeficiente Kappa (κ). Este índice representa la proporción de acuerdos observados respecto del máximo acuerdo posible más allá del azar y no tiene en cuenta el desbalanceo de clases [9]. κ toma valores entre -1 y $+1$ (Tabla 5). Valores de κ cercanos a $+1$

indican un mayor grado de concordancia inter-observador y, más cercano a -1, mayor grado de discordancia [10].

3. Resultados

Los algoritmos de clasificación RF y SVM tienen una precisión global (A_g) de 81 % y 74 % y los coeficientes κ son de 0,8 y 0,7, respectivamente (Tabla 6c).

Tabla 6: Matriz de confusión y evaluación de exactitud a: RF b: exactitud de RF c: Parámetro de presición y *Kappa* d: SVM e: exactitud de SVM f: Parámetro de presición y *Kappa*

(a)

		Predicción				
		M	S	A	U	T
Act.	M	19	22	0	2	43
	S	9	74	0	1	84
	A	0	0	25	2	27
	U	0	0	0	32	32
	T	28	96	25	37	186

(b)

	O	C
M	0,42	0,68
S	0,88	0,78
A	0,92	0,98
U	0,98	0,86

(c)

Param.	Result.
A_g[%]	0,81
κ	0,8

(d)

		Predicción				
		M	S	A	U	T
Act.	M	17	25	0	1	43
	S	21	61	0	2	84
	A	0	0	27	0	27
	U	0	0	0	32	32
	T	38	86	27	35	186

(e)

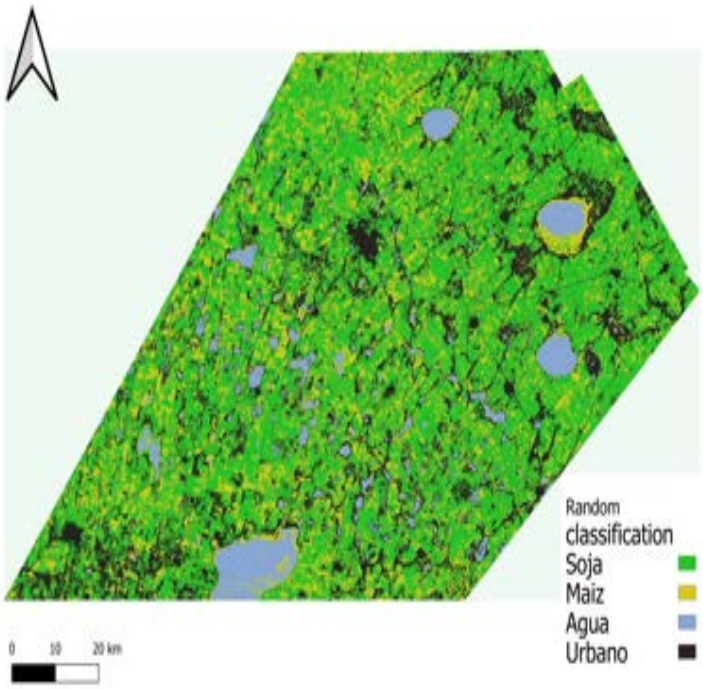
	O	C
M	0,39	0,45
S	0,78	0,75
A	0,98	0,98
U	0,98	0,89

(f)

Param.	Result.
A_g[%]	0,74
κ	0,7

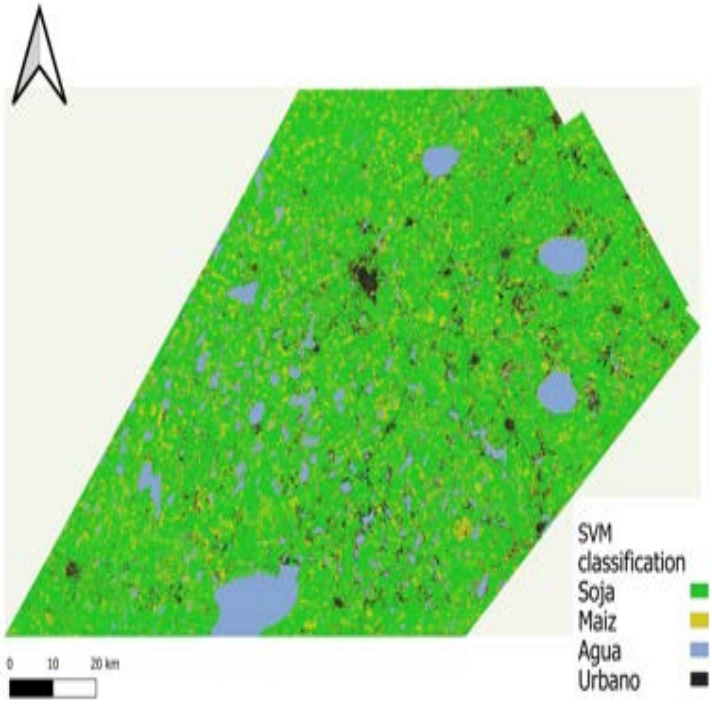
Las matrices de confusión y las mediciones de precisión se detallan en la Tabla 6. Las mayores precisiones (superior a 0,90) son para la cubierta de agua y zonas urbanas en ambos métodos. La soja tiene una precisión de 0,88 en el modelo RF y de 0,78 para SVM.

Los resultados obtenidos son significativos, en la figura 2 se observa el mapeo de las clases con los métodos. En ambas imágenes, se puede ver la correspondencia visual existente.



,

(a) Random Forest



,

(b) Support Vector Machine.

Figura 2: Clasificación de cultivos estivales, cuerpos de agua y zonas urbanas: a: RF. b: SVM.

4. Conclusiones

Es posible diferenciar tipos de cobertura del suelo agrícola, para el área de General López, por medio de clasificaciones automáticas a partir de imágenes Sentinel-2. La clasificación de cultivos estivales, cuerpos de agua y zonas urbanas alcanza índices muy buenos de precisión. De los dos algoritmos evaluados, el de Random Forest proporciona un mejor resultado, tanto en estadios iniciales del cultivo como en avanzadas. En una futura etapa, se prevee relevar más lotes e incluir más categorías, este factor podría mejorar la calidad de los clasificadores. Inclusive, se pretende extrapolar el modelo a otras áreas y analizar su comportamiento.

Referencias

1. Bula, A. O. Importancia de la agricultura en el desarrollo socio-económico,(2020).
2. Willington, E. A., Nolasco, M.,Bocco, M. Clasificación supervisada de suelos de uso agrícola en la zona central de Córdoba (Argentina): comparación de distintos algoritmos sobre imágenes Landsat. In V Congreso Argentino de AgroInformática (CAI)-JAIIO 42 (2013).
3. Vizzotto Cattani, C.E., Mercante, E., Wachholz de Souza, C.H., Costa Wrublack, S.: Desempenho de Algoritmos de Classificação Supervisionada para Imagens dos Satélites RapidEye. En XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR, Foz do Iguaçu, PR. pp 8005–8010. Brasil, (2013).
4. Velasco Cadierno, Raúl.:Aplicación de teledetección para estimación de severidad post-incendio,(2021).
5. Bolsa de Cereales Homepage, https://www.bolsadecereales.com último acceso 25 Jun 2023.
6. Google Earth Engine, http://www.earthengine.google.com último acceso 03 Jul 2023.
7. Ciencia de datos, www.https://www.cienciadedatos.net último acceso 02 Jul 2023.
8. Godoy, Facundo Eduardo. Métodos clásicos de clasificación: comparación y aplicación. BS thesis,2021.
9. Borràs, J., Delegido, J.; Pezzola, A., Pereira, M.,Morassi, G., Camps-Valls, G.: Clasificación de usos del suelo a partir de imágenes Sentinel-2. Revista de Teledetección,(2017).
10. Marini, Mario Fabián.: "Discriminación de cultivos de distinto desarrollo utilizando imágenes satelitales MODIS."GeoFocus. International Review of Geographical Information Science and Technology 131 (2013): 48-60.