EVALUACIÓN DE METODOS MULTIVARIADOS

Jose Manuel Sepulveda Rueda

Taller 7

Edier Aristizábal

Universidad Nacional de Colombia

Facultad de Minas

Cartografía geotécnica

Diciembre 2023

Taller 7

Descripción de la cuenca

La cuenca de la Quebrada Quebradona en el sector occidental del municipio de Ituango, ubicado en el departamento de Antioquia, Colombia, desempeña un papel esencial dentro de la extensa red fluvial del Río Ituango al fungir como uno de sus afluentes.

En su cuenca alta y media, se caracteriza por extensas áreas de bosque que contribuyen significativamente a su biodiversidad. Destacan los siguientes afluentes, como la Quebrada Las Mellizas, la Quebrada Santa Lucía, y la Quebrada Quindío, junto con otros afluentes de menor tamaño, que contribuyen al caudal y la singularidad de esta región.

Además de su relevancia ecológica, la cuenca de la Quebrada Quebradona desempeña un papel crucial en la sustentabilidad de las comunidades locales, que dependen de sus recursos hídricos y disfrutan de los beneficios que brinda a la agricultura y la vida silvestre. Por consiguiente, es imperativo asegurar su conservación y protección, no solo como un valioso ecosistema, sino también como un activo fundamental para el bienestar de aquellos que residen en su entorno.

CARACTERISTICA	DETALLE
ÁREA	76.152497 km^2
PERÍMETRO	46.431991 km
ALTITUD MÁXIMA	3117 msnm
ALTITUD MÍNIMA	1269 msnm
ALTURA PROMEDIO	2252 msnm
LONG AXIAL LARGO	11.8 km
LONG AXIAL ANCHO	11.64 km
PENDIENTE PROMEDIO	30°
LONGITUD DEL CAUCE PRINCIPAL	8.66 km

Tabla 1. Características generales de la cuenca

Métodos Multivariados

Los métodos estadísticos multivariados analizan la relación conjunta entre la variable dependiente (ocurrencia de movimientos en masa) y todas las variables independientes (predictoras) simultáneamente. Uno de estos métodos es la regresión logística, ampliamente utilizado a nivel mundial para evaluar la susceptibilidad a movimientos en masa.

Regresión logística

La regresión logística es un método estadístico multivariado utilizado para predecir la probabilidad de que ocurra un evento binario, es decir, un evento que puede tener dos resultados posibles, como sí (Ocurrencia de MenM) /no (No ocurrencia de MenM), etc. Es especialmente útil cuando la variable que queremos predecir es categórica.

Funciona de la siguiente manera: la regresión logística utiliza una función logística para modelar la relación entre una o más variables independientes (predictoras) y la probabilidad de que ocurra el evento de MenM. Esta función logística transforma la suma ponderada de las variables independientes utilizando una curva en forma de "S", que va de 0 a 1. La salida de esta función representa la probabilidad estimada del evento.

En términos más simples, la regresión logística busca encontrar la mejor combinación de coeficientes para las variables independientes de manera que la curva logística se ajuste de la mejor manera posible a los datos observados. Una vez que se ha ajustado el modelo, podemos usarlo para predecir la probabilidad de que ocurran MenM para nuevos conjuntos de datos.

Python

Nota: para visualizar el código de Python de cada una de las figuras, dirigirse al documento anexo a este en la carpeta.

Se usa la librería statsmodels para mirar un resumen de los resultados y métricas del modelo

Logit Regression Results

Dep. Variable:	inventario	No. Observations:	484040
Model:	Logit	Df Residuals:	484030
Method:	MLE	Df Model:	9
Date:	Mon, 04 Dec 2023	Pseudo R-squ.:	0.03839
Time:	15:42:34	Log-Likelihood:	-72323.
converged:	False	LL-Null:	-75210.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-16.2717	586.859	-0.028	0.978	-1166.495	1133.951
C(geologia)[T.1.0]	14.7331	586.859	0.025	0.980	-1135.490	1164.956
C(geologia)[T.510.0]	15.2805	586.859	0.026	0.979	-1134.942	1165.503
C(geologia)[T.765.0]	14.7425	586.859	0.025	0.980	-1135.480	1164.965
C(geologia)[T.1020.0]	-12.5388	1.02e+05	-0.000	1.000	-1.99e+05	1.99e+05
pendiente	0.2069	0.008	26.910	0.000	0.192	0.222
dem	-0.0011	2.14e-05	-49.595	0.000	-0.001	-0.001
Curvatura	0.0059	0.004	1.646	0.100	-0.001	0.013
flujo_acum	0.0063	0.001	5.367	0.000	0.004	0.009
aspecto	-0.2552	0.008	-31.399	0.000	-0.271	-0.239

Figura 1. Resumen del modelo

Se reconstruye la cuenca con los valores de susceptibilidad. Para eso utilizaremos como máscara el mapa de pendiente.

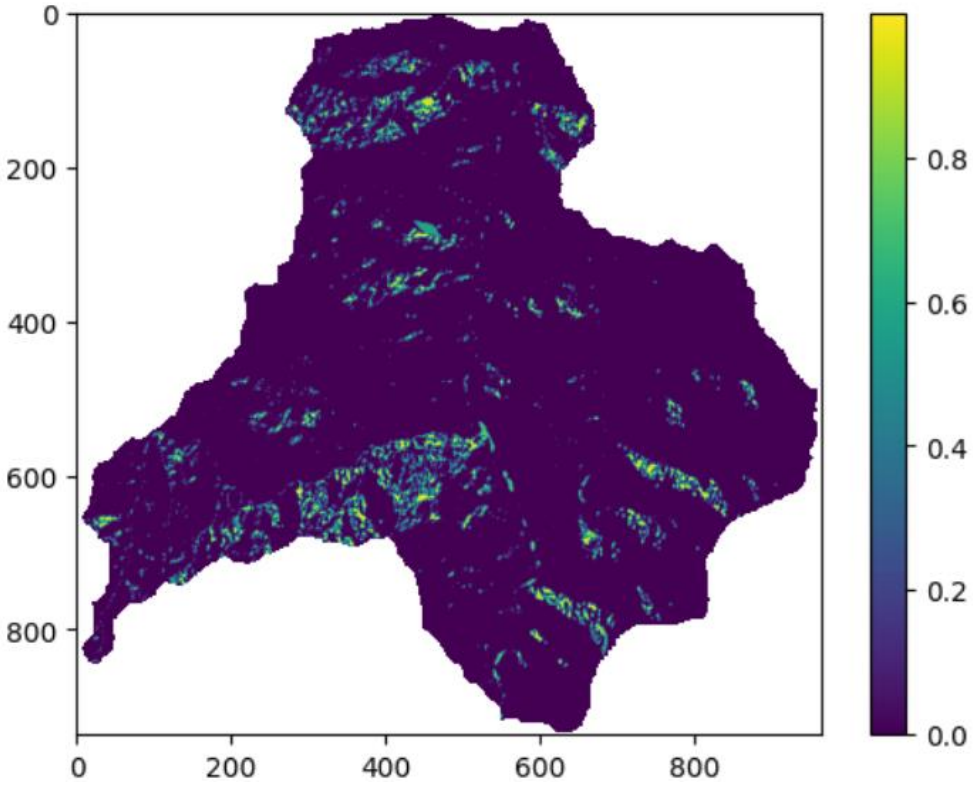

Figura 2. Cuenca con los valores de susceptibilidad RL

Evaluación del modelo RL

la evaluación del modelo debe llevarse a cabo en dos aspectos fundamentales: el rendimiento y la capacidad predictiva. La evaluación más desafiante recae en la capacidad predictiva, dado que solo la ocurrencia de eventos futuros puede determinar verdaderamente la efectividad del modelo en este sentido. Sin embargo, es crucial realizar evaluaciones tanto del rendimiento como de la capacidad predictiva durante la construcción del modelo.

Validación cruzada

```
Tamaño de variables de entrenamiento: (387232, 6)
Tamaño de labels de entrenamiento: (387232,)
Tamaño de variables de validación: (96808, 6)
Tamaño de labels de validación: (96808,)
```

Figura 3. Validación de las variables RL

Desempeño del modelo

Curva ROC

La matriz de contingencia se puede representar en un espacio bidimensional que contrasta la tasa de verdaderos positivos o Sensibilidad en el eje y con la tasa de falsos positivos en el eje x. Este enfoque nos facilita la comparación del rendimiento entre diversos modelos. El modelo con el rendimiento teóricamente óptimo se sitúa en la esquina superior derecha del gráfico, indicando un 100% de aciertos y un 0% de errores. La distancia euclidiana a este punto representa la Distancia a la Clasificación Perfecta (r). En consecuencia, los modelos con una menor distancia exhiben un rendimiento superior.

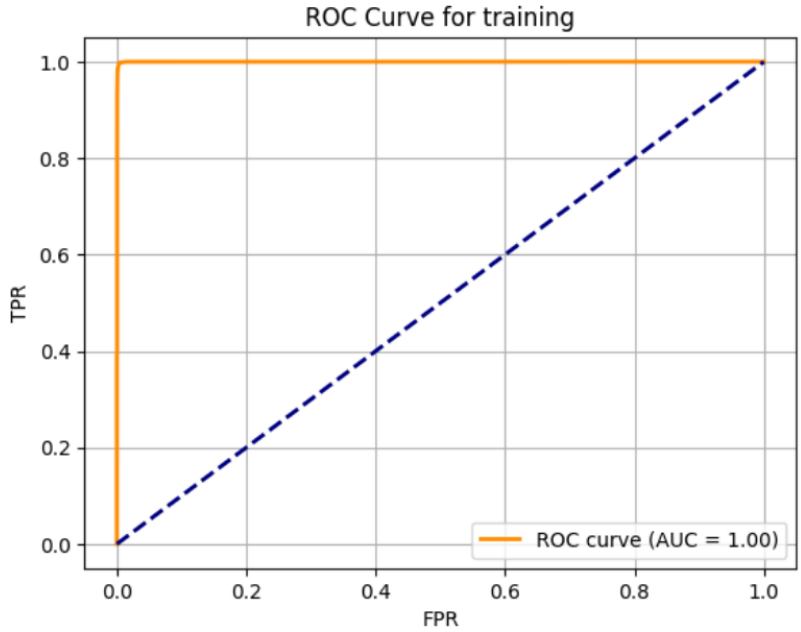

Figura 4. Curva ROC para el entrenamiento en RL

Curva ROC para predicción

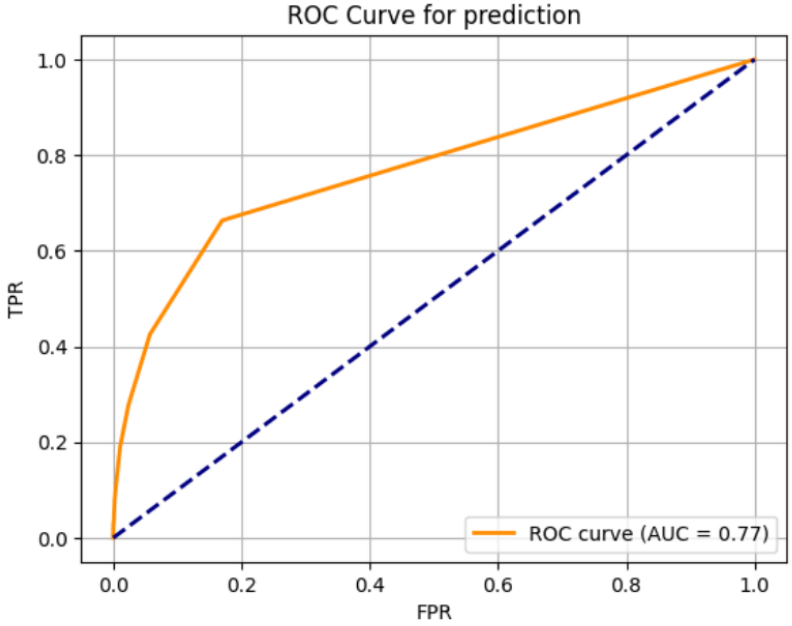

Figura 5. Curva ROC para la predicción en RL

valores del index de susceptibilidad

	fpr	tpr	1-fpr	tf	threshold
0	0.000000	0.000000	1.000000	-1.000000	2.000000
1	0.000032	0.007332	0.999968	-0.992636	1.000000
2	0.000075	0.016356	0.999925	-0.983568	0.900000
3	0.000097	0.016356	0.999903	-0.983547	0.819030
4	0.000300	0.034405	0.999700	-0.965295	0.800000
5	0.000322	0.034405	0.999678	-0.965273	0.762683
6	0.000343	0.034405	0.999657	-0.965252	0.745040
7	0.000622	0.035533	0.999378	-0.963845	0.704269
8	0.000633	0.035533	0.999367	-0.963834	0.702345
9	0.001062	0.057248	0.998938	-0.941691	0.700000
10	0.001104	0.057530	0.998896	-0.941366	0.680886
11	0.001137	0.057530	0.998863	-0.941334	0.658721
12	0.002198	0.090243	0.997802	-0.907559	0.600000
13	0.002220	0.090243	0.997780	-0.907538	0.552728
14	0.004686	0.125776	0.995314	-0.869539	0.500000

Figura 6. Index de susceptibilidad utilizados para cada punto de la curva en RL

Para estimar el punto más cercano a la clasificación perfecta (esquina superior izquierda de ROC), el cual corresponde al umbral que presenta la mejor matriz de confusión se puede realizar lo siguiente

	fpr	tpr	1-fpr	tf	threshold
25	0.166477	0.656232	0.833523	-0.17729	0.084273

Figura 7. Utilizar la función argsort curva en RL

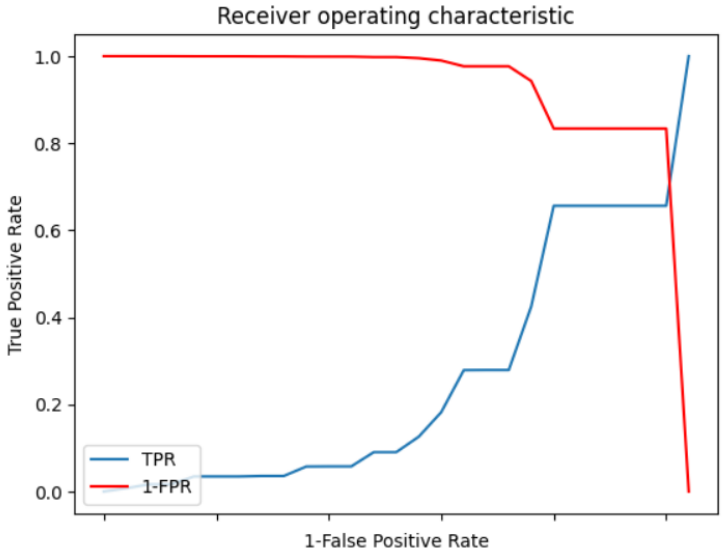

Figura 8. Grafica de TPR y 1-FPR. El intercepto de dichas gráficas corresponde al umbral buscado en RL

Otra representación ampliamente empleada, especialmente en situaciones de conjuntos de datos desbalanceados, como ocurre en este caso, es la gráfica de precisión y recuperación (precision-recall).

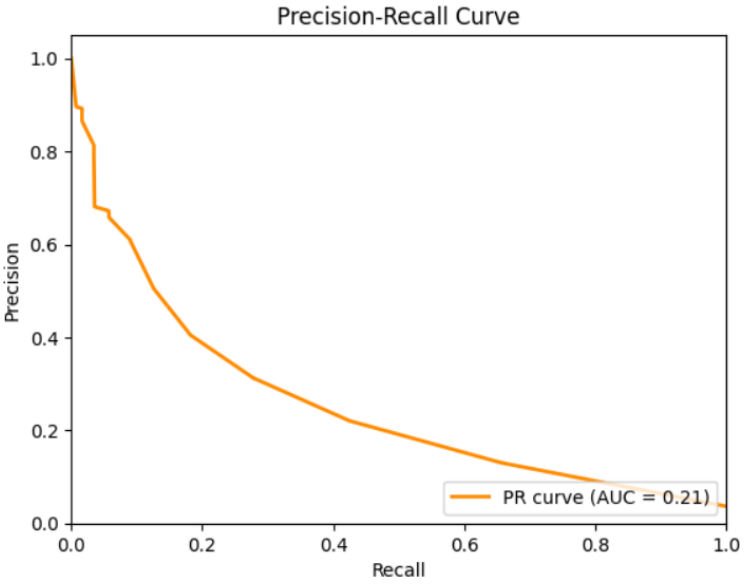

Figura 9. Grafica precision-recall en RL

Mapa de susceptibilidad final

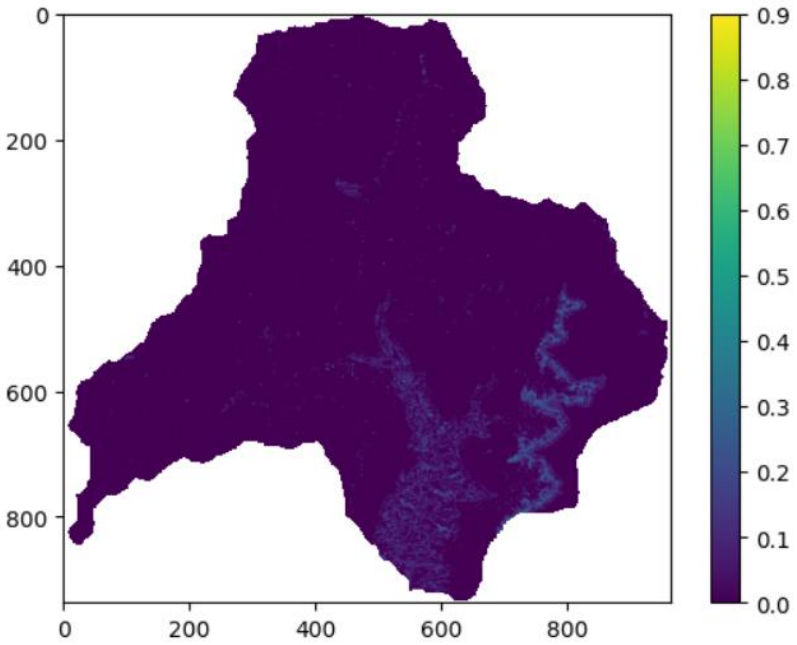

Figura 10. Mapa de susceptibilidad final del método de regresión lineal

Nota: Se realizo otras 3 evaluaciones de métodos multivariados, Análisis Discriminante Lineal, Support Vector Machine y Redes Neuronales, para ver el código a detalle diríjase al documento anexo de esta carpeta

Los mapas de susceptibilidad dados por estos métodos son:

Análisis Discriminante Lineal

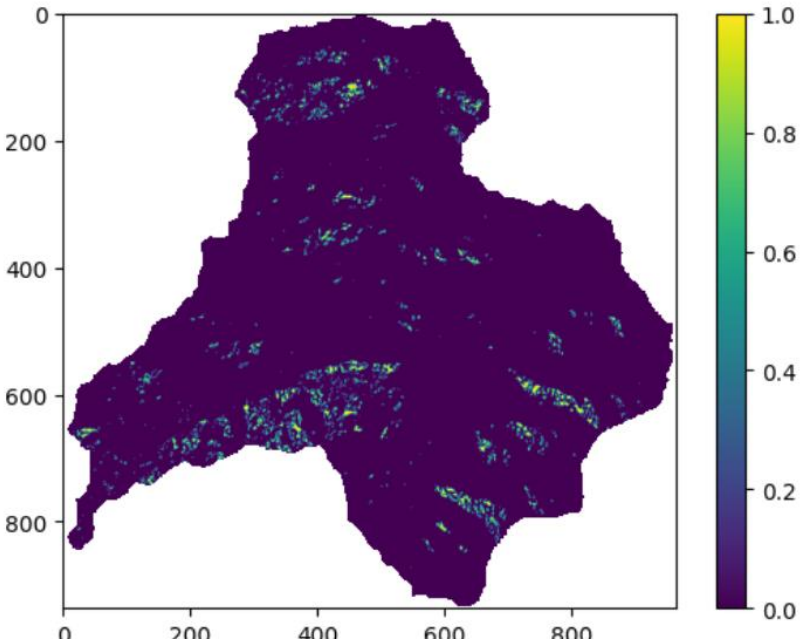

Figura 11. Cuenca con los valores de susceptibilidad con el método de análisis discriminante lineal

Se evalúa el método y se muestra en siguiente tabla lo más relevante.

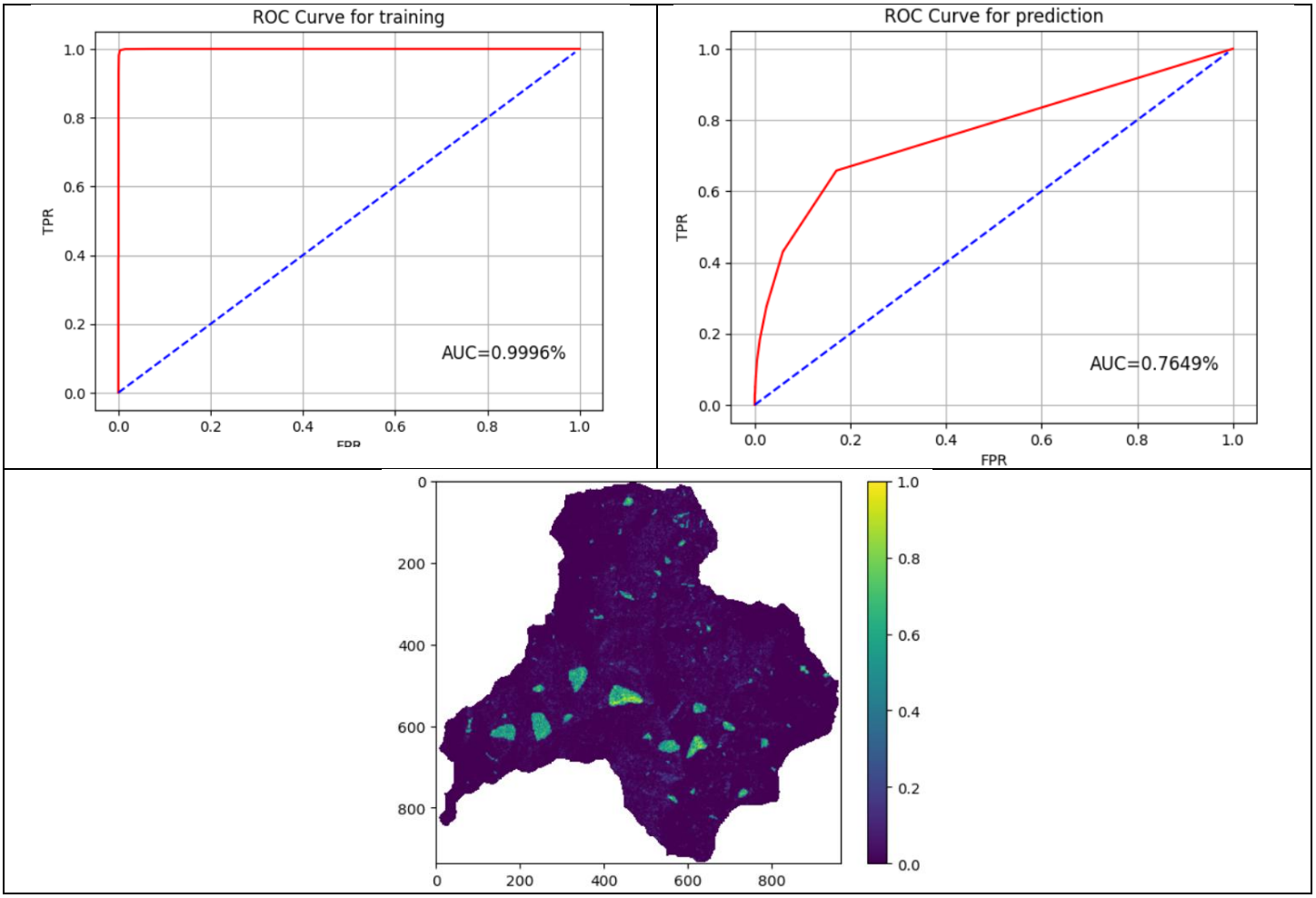

Figura 12. Evaluación del método de análisis discriminante lineal

Support vector machine

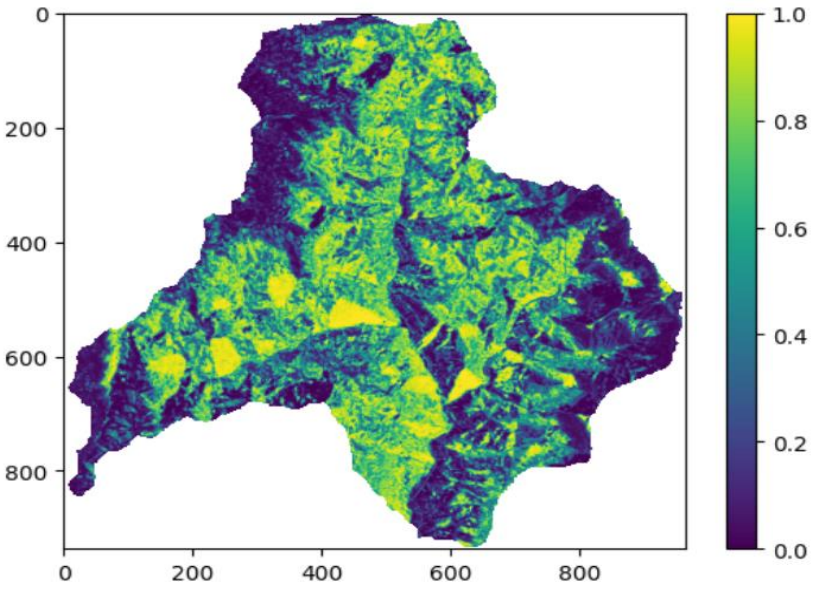

Figura 13. Cuenca con los valores de susceptibilidad con el método de support vector machine

Se evalúa el método y se muestra en siguiente tabla lo más relevante.

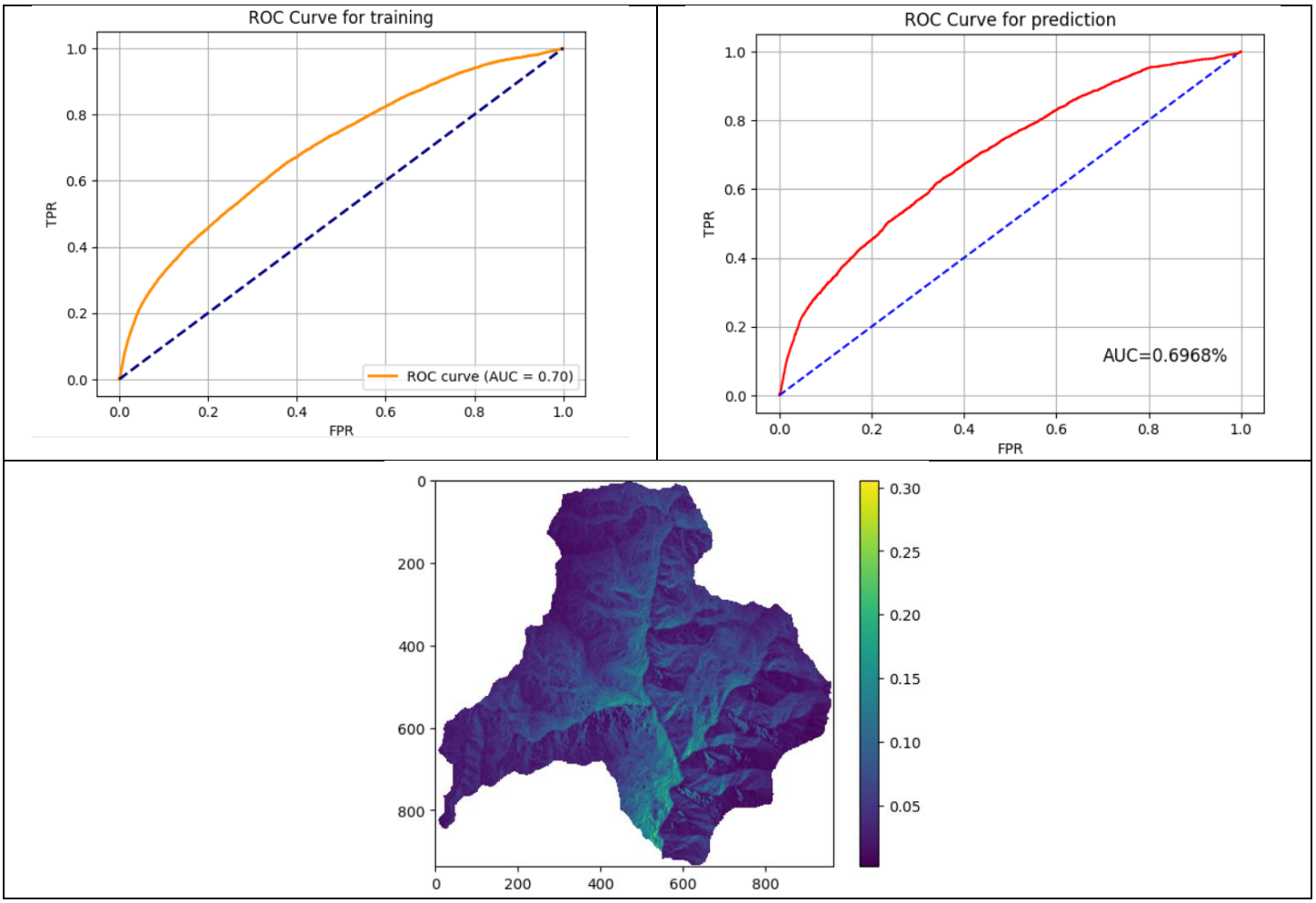

Figura 14. Evaluación del método de support vector machine

Redes Neuronales

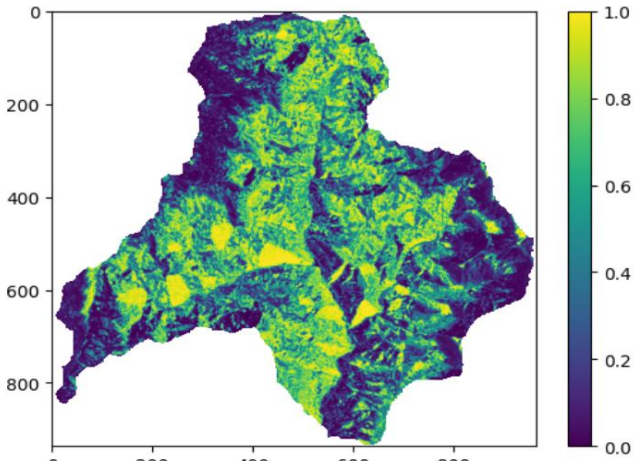

Figura 15. Cuenca con los valores de susceptibilidad con el método de redes neuronales

Se evalúa el método y se muestra en siguiente tabla lo más relevante

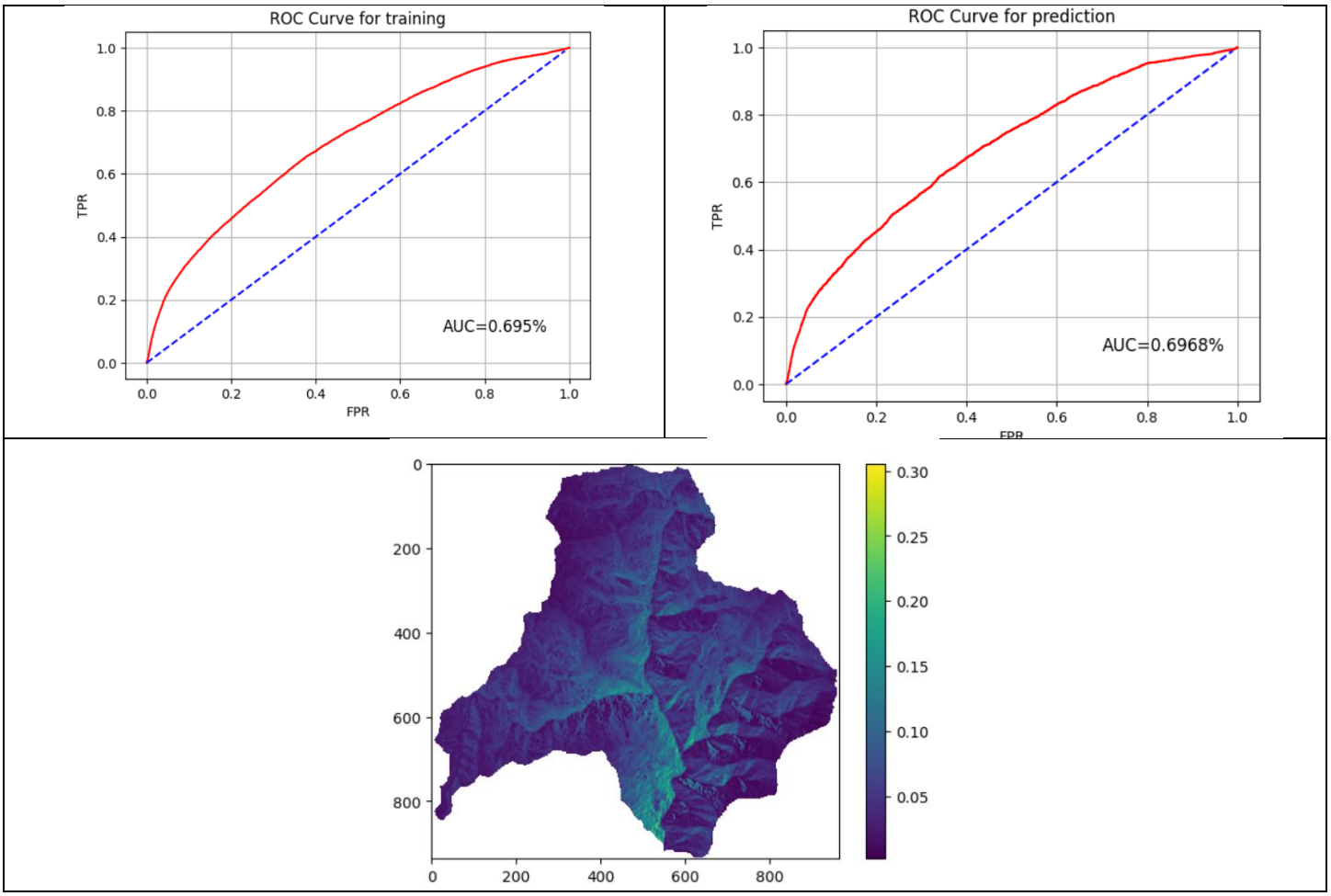

Figura 16. Evaluación del método de redes neuronales