Estimación del Riesgo de Deslizamientos Mediante Algoritmos de Aprendizaje Automático (Vía Calacalí-Nanegalito)

Estimation of the Risk of Landslides Using Automatic Learning Algorithms (Way Calacalí-Nanegalito)

Bustos D.[1]; Estrada S.[2]; Soria G.[3]; Mejía C.[4]

[1] Universidad Central del Ecuador, Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental. Carrera de Geología. Quito, Ecuador
e-mail: dsbustos@uce.edu.ec

[2] Universidad Central del Ecuador, Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental. Carrera de Geología. Quito, Ecuador
e-mail: csestrada@uce.edu.ec

[3] Universidad Central del Ecuador, Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental. Carrera de Geología. Quito, Ecuador
e-mail: gisoria@uce.edu.ec

[4] Universidad Central del Ecuador, Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental. Carrera de Geología. Quito, Ecuador
e-mail: cimejia@uce.edu.ec

RESUMEN

En la presente investigación se han aplicado diferentes algoritmos de Machine Learning: 1. Regresión Logística; 2. Máquina de Vector de Soporte (SVM), 3. Redes Neuronales, el objetivo es desarrollar un código que genere predicciones de futuros deslizamientos y así poder generar mapas de susceptibilidad a deslizamientos en la vía Calacalí – Nanegalito, esto se logró utilizando un inventario de deslizamientos que se ha realizado mapeando en campo a lo largo de la vía [1]. Se han extraído diferentes variables que son influyentes en la generación de deslizamientos, estas son: la litología, elevación, aspecto, perfil de curvatura, plano de curvatura, TWI, Índice de vegetación, Magnitud sísmica y susceptibilidad a deslizamientos. Todas las variables fueron analizadas estadísticamente para evaluar la distribución de datos y analizar sesgos que puedan afectar a las predicciones junto con una normalización de todos los datos. Además, se realizó la preparación de la base de datos de entrenamiento y prueba (training 80% – test 20%) esto con datos positivos 50% (deslizamiento) y negativos 50% (no deslizamiento) equilibrados. Como resultado de la aplicación de estos 3 modelos de Machine Learning se obtuvo 3 diferentes mapas de susceptibilidad de deslizamientos los cuales han sido comparados con el Método de Mora- Varhson y se llegó a la conclusión que el modelo de redes neuronales presenta una exactitud de 99,3% por lo que estadísticamente es el modelo más preciso, sin embargo el modelo de regresión logística obtuve mejores resultados comparados con las zonas de mayor susceptibilidad que se generaron con el mapa de Método de Mora- Varhson.

Palabras clave: Machine Learning, Regresión Logística, Máquina de soporte de vector, Redes Neuronales, Susceptibilidad, Deslizamiento.

ABSTRACT

In this research, different Machine Learning algorithms have been applied: 1. Logistic Regression; 2. Support Vector Machine (SVM), 3. Neural Networks, the objective is to develop a code that generates predictions of future landslides and thus to be able to generate maps of susceptibility to landslides on the Calacalí - Nanegalito road, this was achieved using a inventory of landslides that has been carried out mapping in the field along the road [1]. Different variables have been extracted that are influential in the generation of landslides, these are: lithology, elevation, aspect, curvature profile, plane of curvature, TWI, Vegetation index, seismic magnitude and susceptibility to landslides. All the variables were statistically analyzed to evaluate the data distribution and analyze biases that could affect the predictions together with a normalization of all the data. In addition, the preparation of the training and test database (training 80% - test 20%) was carried out with balanced positive 50% (slip) and negative 50% (no slip) data. As a result of the application of these 3 Machine Learning models, 3 different landslide susceptibility maps were obtained which have been compared with the Mora-Varhson Method and it was concluded that the neural network model has an accuracy of 99, 3%, so statistically it is the most accurate model, however the linear regression model obtained better results compared to the areas of greater susceptibility that were generated with the Mora-Varhson method map.

Keywords: Machine Learning, Logistic Regression, Support Vector machine, Neural Network, Susceptibility, Landslide.

1. INTRODUCCIÓN

Los deslizamientos de tierra son fenómenos complejos que pueden originarse de manera natural, bajo la influencia directa de la gravedad, o por acción del ser humano, por obras de infraestructura construidas en lugares inadecuados. En ambos casos, hay un desequilibrio de esfuerzos en los taludes provocando un rápido movimiento de una porción considerable de escombros, suelo o rocas hacia la parte inferior de las pendientes. Estos fenómenos amenazan al mismo ser humano y la infraestructura, causando varias muertes y daños materiales.

El *Aprendizaje Automático (Machine Learning*) es una técnica aprovechada recientemente y considerada como una alternativa conveniente para el tratamiento de este tipo de fenómenos. Básicamente, a partir de datos ya existentes como un inventario de deslizamientos anteriores y comprobados en campo, permite establecer modelos matemáticos con las variables necesarias para predecir el lugar donde podría darse un posible deslizamiento.

Nuestro caso de estudio es la vía Calacalí-Nanegalito, una de las principales rutas que conectan las regiones Sierra y Costa del Ecuador. En los últimos años, son varios los deslizamientos en este sector, mismos que han ocasionado pérdidas humanas y económicas. Por tanto, se requieren estudios técnicos con el propósito de mitigar los efectos de este fenómeno.

La parroquia de Calacalí se encuentra a 17 km al norte de Quito, cerca de la Ciudad Mitad del Mundo (Figura 1). La vía Calacalí-Nanegalito es de gran importancia ya que conecta el norte de Quito con la costa ecuatoriana siendo muy transitada por los ciudadanos, sin embargo, en este sector se han producido gran cantidad de deslizamientos. El presente estudio analiza un tramo de 17.5 km y un área de influencia de 320m alrededor de dicha vía.

1.1 UBICACIÓN:

La parroquia de Calacalí se encuentra a 17 km al norte de Quito, cerca de la Ciudad Mitad del Mundo, el tramo analizado corresponde 17.5 km de la vía Calacalí Nanegalito, esta vía es gran importancia ya que conecta el norte de Quito con la costa ecuatoriana siendo muy transitada por los ciudadanos, en este sector se han producido gran cantidad de deslizamientos alrededor de la vía Calacalí Nanegalito es por esto que nuestro estudio se centró en 220m alrededor de dicha vía.

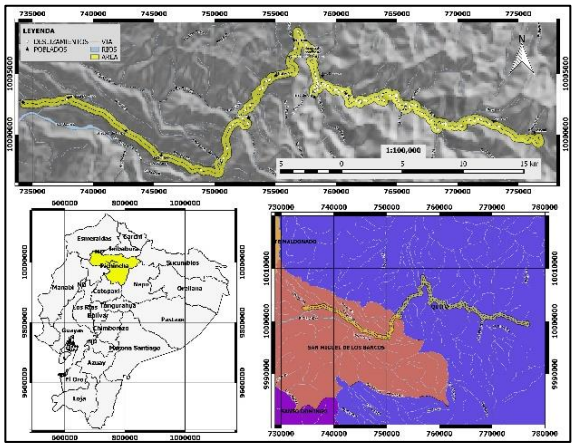

Figura 1. Vía Calacalí-Nanegalito.

2. METODOLOGÍA

Se representa por medio del siguiente diagrama de flujo (Figura 2):

2.1 DATASET

El principal insumo para el trabajo con Machine Learning es un conjunto de datos (*dataset*) suficiente y correcto. Para tal fin, se utilizó un inventario deslizamientos a lo largo de la vía, en donde se han identificado las coordenadas geográficas de cada deslizamiento, obteniendo un total de 78 puntos cartografiados.

Esta información es tratada en el software QGIS versión 3.12 para la creación de un área de influencia (*buffer*) de 320 m a cada lado de la vía esta distancia se determinó tomando en cuenta la longitud de las laderas que atraviesa la vía. Posteriormente, desde el sitio Web de Opentopography (https://www.opentopography .org), se descargó el modelo digital de elevación (DEM) de la zona de interés con un tamaño de celda (píxel) de 20x20 metros.

2.1.1Variables

Para la construcción del dataset y aprovechando el DEM, se elaboraron mapas temáticos de litología, elevación, pendiente, aspecto, perfil de curvatura, plano de curvatura, vegetación, TWI (Índice topográfico de humedad), humedad. De cada mapa (Figuras 3 a 12) se extrae la información de cada píxel con el propósito de obtener datos confiables.

Por lo tanto, se identificaron 10 variables independientes (factores de ocurrencia del fenómeno) y una variable dependiente (susceptibilidad de deslizamiento). La Tabla 1 presenta todas las variables consideradas.

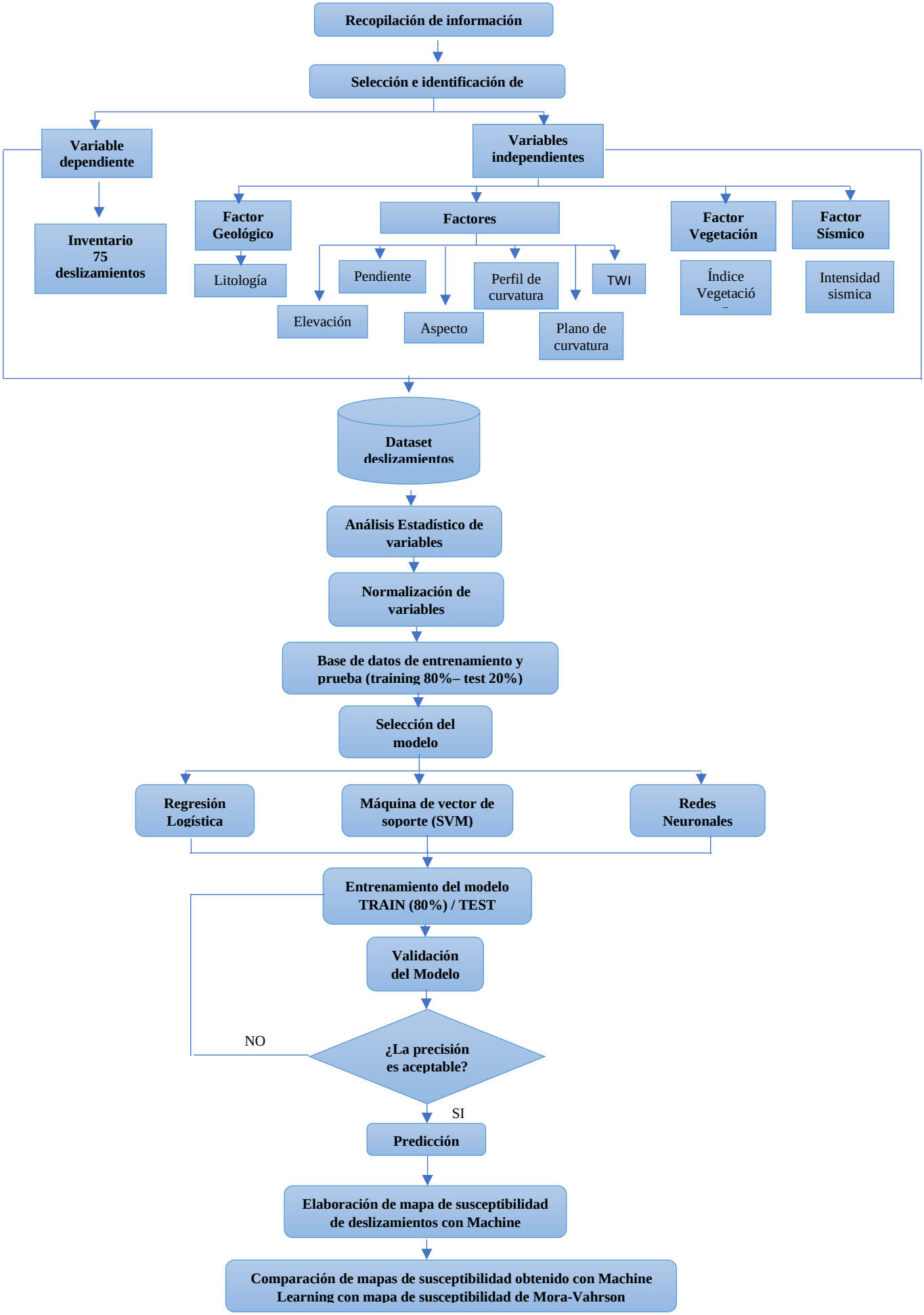

Figura 2. Metodología aplicada para predicción de deslizamientos [2].

NOMBRE	DESCRIPCIÓN	TIPO	UNIDADES DE MEDIDA	ML	AREA DE ESTUDIO
SUSCEPTIBILIDAD	Categoría o nivel de riesgo de deslizamiento	Cualitativa	ADIMENSIONAL	DEPENDIENTE	Gestión de Riesgos
ELEVACION	Representa los valores de altura con respecto al nivel del mar	Cuantitativa	METROS	INDEPENDIENTE	Geomorfológica
ASPECTO	Dirección de la pendiente de cada píxel	Cuantitativa	GRADOS SEXAGESIMALES	INDEPENDIENTE	Geomorfológica
PENDIENTE	Representa la diferencia de gradiente entre dos formas de relieve	Cuantitativa	PORCENTAJE	INDEPENDIENTE	Geomorfológica
LITOLOGIA	Tipo de roca	Cualitativa	ADIMENSIONAL	INDEPENDIENTE	Geología
HUMEDAD	Concentración del contenido de agua en un sustrato	Cuantitativa	PORCENTAJE	INDEPENDIENTE	Hidrogeología
TWI (Índice topográfico de humedad)	Relativa concentración de humedad en un sustrato	Cuantitativa	PORCENTAJE	INDEPENDIENTE	Geomorfológica
Intensidad Sisimica	Escala de Mercalli en función de las afectaciones	Cuantitativa	ADIMENSIONAL	INDEPENDIENTE	Sísmica
PLANO DE CURVATURA	Perpendicularidad a la dirección de la pendiente	Cuantitativa	ADIMENSIONAL	INDEPENDIENTE	Geomorfológica
PERFIL DE CURVATURA	Es la paralela a la pendiente e indica la dirección de pendiente máxima	Cuantitativa	ADIMENSIONAL	INDEPENDIENTE	Geomorfológica
VEGETACION	Áreas de uso de suelo	Cualitativa	ADIMENSIONAL	INDEPENDIENTE	Uso de suelo

Tabla 1. Resumen de las variables dependientes e independientes.

2.1.2 Mapas de variables

Mapa Litológico

El mapa litológico (Figura 3) se realizó considerando la información de la carta geológica de Pacto y Quito a escala 1:100.000. La clasificación está de acuerdo con la litología correspondiente a cada formación.

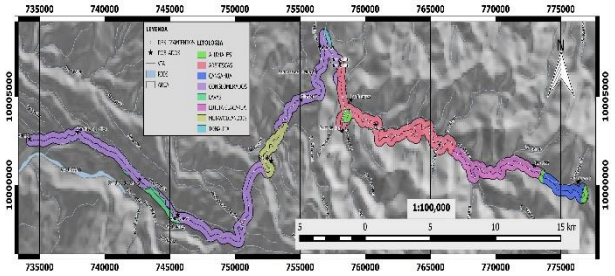

Figura 3. Mapa litológico.

Mapa de Elevación

El mapa de elevación (Figura 4) se lo determinó mediante la adquisición de un Modelo de Elevación Digital (DEM) extraídos de fotos satelitales con una resolución aproximada de 30 metros del tamaño de pixel de la zona de estudio, con alturas entre 1360 a 2428 msnm.

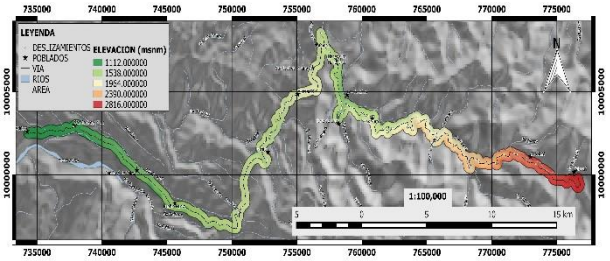

Figura 4. Mapa de elevación.

Mapa de Pendientes

El mapa de pendientes (Figura 5) y sus derivados (aspecto, perfil de curvatura, plano de curvatura, y TWI) son extraídos a partir del DEM, con un tamaño de celda de 20x20 metros, donde se aprecian pendientes entre 0 a 65º.

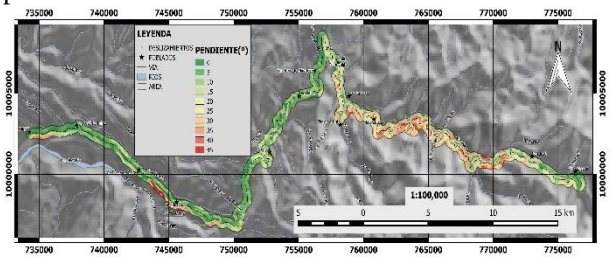

Figura 5. Mapa de pendientes.

Mapa de Plano de Curvatura

El plano de curvatura representa la perpendicular a la dirección de la pendiente máxima. La curvatura del plano se relaciona con la convergencia y divergencia de la corriente por una superficie [8]. Un valor positivo indica que la superficie es lateralmente convexa en esa celda. Un plano negativo indica que la superficie es lateralmente cóncava en esa celda. Un valor de cero indica que la superficie es lineal. El plano de curvatura es muy influyente en deslizamientos en valores cercanos a cero donde indica que una superficie es lineal y un fluido puede afectar a una extensa área de una determinada litología siendo esta propensa a la erosión e infiltración de agua provocando una alta probabilidad de ocurrencia a deslizamientos (Figura 6)

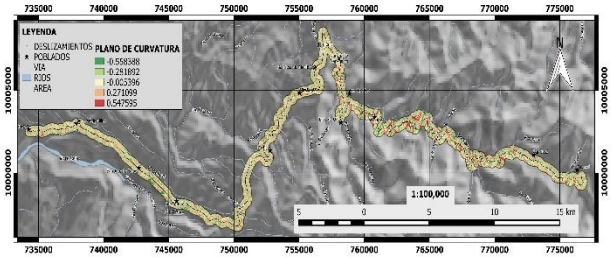

Figura 6. Mapa de Plano de Curvatura.

Mapa de Aspecto

En este caso el mapa de aspecto nos indica la orientación de la inclinación de la pendiente, este es favorable para identificar con coherencia la dirección de deslizamiento; a lo largo de la zona de estudio se puede encontrar pendientes orientadas 0 a 360º de azimut (Figura 7)

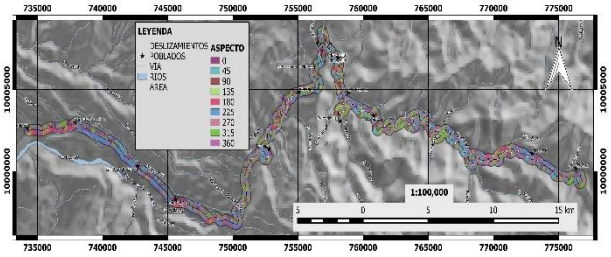

Figura 7. Mapa de Aspecto.

Mapa de Perfil de Curvatura

El perfil de curvatura está relacionado con la aceleración y desaceleración de las corrientes por la superficie, siendo así muy influyente en litologías que tienen un perfil negativo donde hay una desaceleración de fluidos (agua) permitiendo la infiltración de esta, provocando así una desestabilización de un cuerpo de tierra para provocar un deslizamiento. (Figura 8)

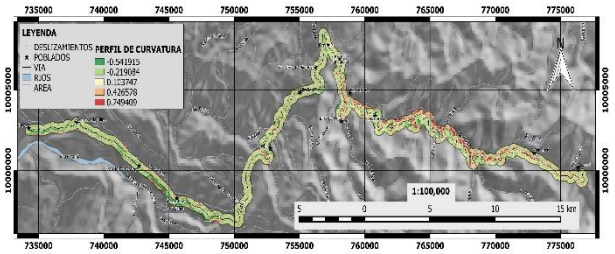

Figura 8. Mapa de Perfil de Curvatura.

Mapa de Índice Topográfico de Humedad (TWI)

El Índice Topográfico de Humedad (TWI) identifica las zonas de mayor potencial de acumulación de humedad. El resultado será un grupo de zonas territoriales de gran potencial de recepción de agua pudiendo delimitarlas como zonas potenciales de surgencia de humedales que pueden ocasionar una alta probabilidad para la ocurrencia de deslizamientos. (Figura 9)

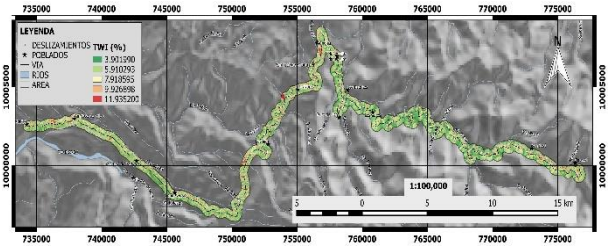

Figura 9. Mapa de TWI.

Mapa de intensidades sísmicas

El mapa de aceleraciones sísmicas se determinó mediante sismos anteriores que han afectado a la zona de estudio, son muy pocas las localidades que presentan este problema. Por lo tanto, se le ha clasificado en función al mayor evento sísmico producido en el país, el terremoto de Riobamba, el 4 de febrero de 1797. (Figura 10)

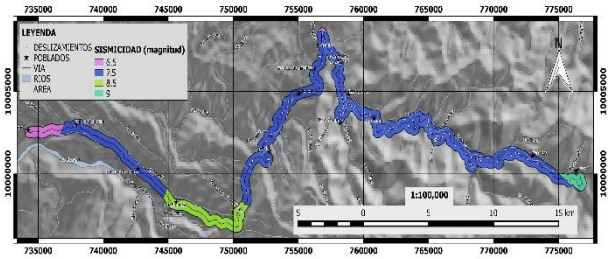

Figura 10. Mapa de Intensidades Sísmicas.

Mapa de Vegetación

El mapa de vegetación (Figura 11) se lo obtuvo mediante información disponible en la base de datos de MAGAP, en la cual categoriza el uso de suelo de todo el país, para este caso se cortó mediante QGIS la información pertinente para la zona de estudio.

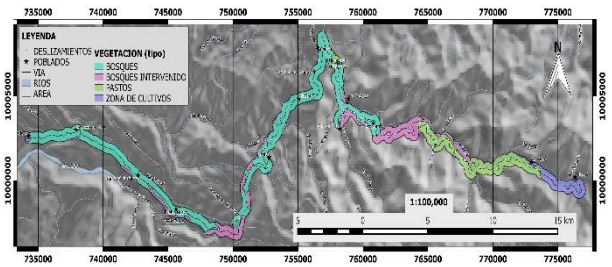

Figura 11. Mapa de Vegetación.

Mapa de Humedad

El mapa de humedad (Figura 12) muestra la humedad relativa en porcentaje (%) con la ayuda de una escala de color. Este índice de humedad se determina como el cociente entre la precipitación y la evapotranspiración potencial, es un indicador representativo del déficit o excedente de los recursos hídricos necesarios para el desarrollo vegetal y, por tanto, de las condiciones de humedad.

Uno de los factores más importantes y que desencadena deslizamientos es la humedad, ya que por la infiltración de agua hace que aumente la presión intersticial provocando fuerzas desestabilizadoras dentro de una masa de suelo o roca

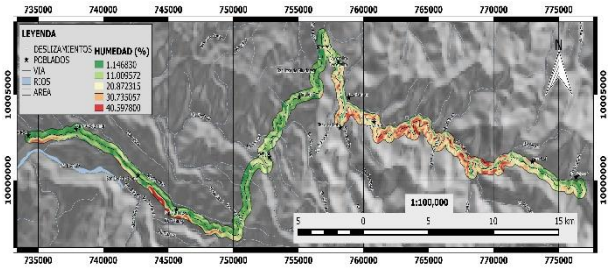

Figura 12. Mapa de Humedad.

2.2. PREPROCESAMIENTO

2.2.1 Análisis estadístico exploratorio

Se ha realizado estadística descriptiva de 110 915 datos de las 10 variables independientes y la variable dependiente, con el objetivo determinar la distribución de los datos y saber si existen valores atípicos o sesgos en los datos que puedan influir en el aprendizaje autónomo de los algoritmos. Se han obtenido los siguientes resultados:

Variable	Unidad	Min.	1er Cuartil	Mediana	Media	3er Cuartil	Max.	Sesgo	Desviación Estándar
Elevación	m.s.n.m	908	1522	1729	1803	2014	2937	Simétrica	442.89
Perfil de curvatura	adimensional	-2.770	-0.187	0.00	0.0318	0.24	2.4637	simétrica	0.3105
Plano de curvatura	adimensional	-2.00	0.00	0.00	0.0021	0.00	1.00	asimétrico a la derecha	0.2369
TWI	Porcentaje	2.764	5.491	6.266	6.557	7.229	19.476	asimétrico a la derecha	1.729
Humedad	Porcentaje	0.000	6.574	13.422	16.155	24.821	63.482	asimétrico a la derecha	11.28
Aspecto	Grados Sexagesimales	-1.00	61.39	210.96	185.29	294.44	359.59	Bimodal asimétrico	118.82
Intensidad Sísmica	Adimensional	6.50	7.50	7.50	7.62	7.50	9.00	asimétrico a la derecha	0.4674
Pendiente	Porcentaje	0.000	6.781	14.103	16.317	24.848	58.552	asimétrico a la derecha	11.170
Susceptibilidad	Adimensional	0.00	0.00	0.00	0.0091	0.00	1.00	asimétrico a la derecha	0.499

Tabla 2. Resumen del análisis de estadística básica de los datos de las variables cuantitativas.

Susceptibilidad de deslizamientos

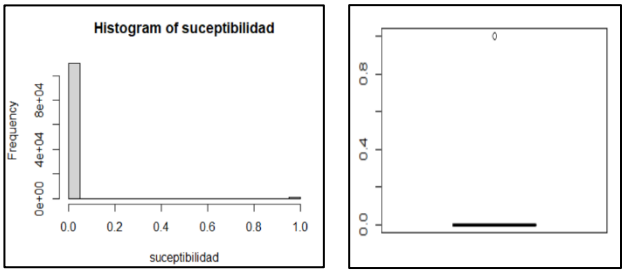

Figura 13. Análisis de estadística básica de los datos de la variable susceptibilidad de deslizamientos.

Balance de datos

Claramente existe un desequilibrio de datos por lo que se utilizó los datos de entrenamiento con valores más parejos para datos positivos como negativos.

En cuanto a la variable dependiente que es el factor a predecir deben estar balanceadas tanto con datos positivos (deslizamiento) como negativos (no deslizamiento) para que el algoritmo no genere sesgos en el aprendizaje.

Se han determinado cerca de 2068 datos de entrenamiento, los valores positivos de la variable dependiente (susceptibilidad a deslizamientos) fueron determinados en base al inventario de deslizamientos que se cuenta de la vía Calacalí – Nanegalito obteniendo 1020 datos positivos y los datos negativos se obtienen considerando pendientes menores a 4%, vegetación tipo bosque y litologías tipo lavas y metavolcánicos obteniendo 1048 datos negativos, obteniendo una balance de datos positivos y negativos.

Normalización de variables

Es importante debido a las diferentes escalas y unidades de medida en las que se encuentran los valores de las diversas variables. Caso contrario, podrían generar sesgos al incorporarse al algoritmo de aprendizaje automático, afectando los resultados de la predicción.

Para normalizar los datos se utilizó la librería RSNNS incluida en el software R Studio, en el cual el tipo de normalización es 0_1 en el cual escala todos los valores de las variables entre 0 y 1

2.3 División de datos: Training-test

La técnica de Machine Learning necesita un porcentaje de los datos para realizar el proceso de entrenamiento (training data).

Por otro lado, los datos de prueba o test son los que ayudan a evaluar el grado de acierto que tiene el modelo de machine learning aprendido en base a los datos de entrenamiento. Los porcentajes de datos que se han utilizado son del 80% para datos de entrenamiento y el 20% restante para datos de test.

2.4 Creación del modelo

2.4.1 Regresión logística

La Regresión Logística lleva el nombre de la función utilizada en el núcleo del método, la función logística es también llamada función Sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1.

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO. Por su parte si el resultado es 0.75.

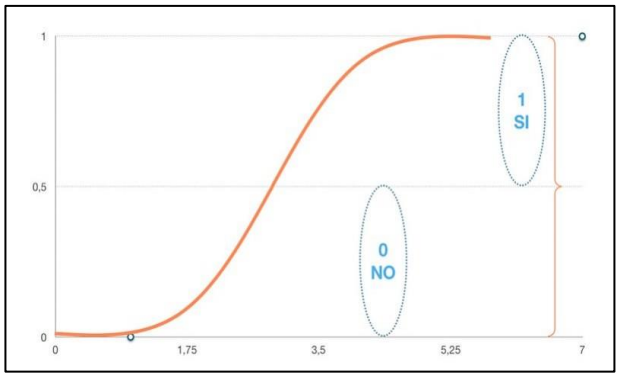

Figura 14. Gráfica ejemplo del modelo de regresión logística [3].

El método de regresión logística permite estimar una probabilidad de una variable cualitativa (deslizamientos) en función de variables cuantitativas (pendiente, plano de curvatura, perfil de curvatura, etc.). Para lo cual todas dichas variables se suman obteniendo una predicción de deslizamientos. Los coeficientes positivos de las variables muestran la importancia de cada parámetro en el momento de realizar la estimación. De esta forma se observa que la humedad (7.81) y pendiente (3.47) son los que más influencian. Mientras que presentan menor relevancia el aspecto y plano de curvatura (coeficientes menores a 0).

2.4.2 Modelo de máquina de soporte vectorial (SVM)

La máquina de soporte vectorial (SVM) son modelos de aprendizaje supervisado, se utiliza principalmente en problemas de clasificación. [4]

Una máquina de vectores de soporte (SVM) es un clasificador discriminativo definido formalmente por un hiperplano separador (Figura 15). En otras palabras, dados los datos de entrenamiento etiquetados (aprendizaje supervisado), el algoritmo genera un hiperplano óptimo que categoriza nuevos ejemplos.

Para realizar una clasificación binaria se utiliza en el método SVM se utilizan hiperplanos el cual permite clasificar a que grupo pertenece cada observación en función de sus variables independientes. El hiperplano representa una recta cuya ecuación es:

$$\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp<0, \text{ si } yi=-1 \quad (1)$$

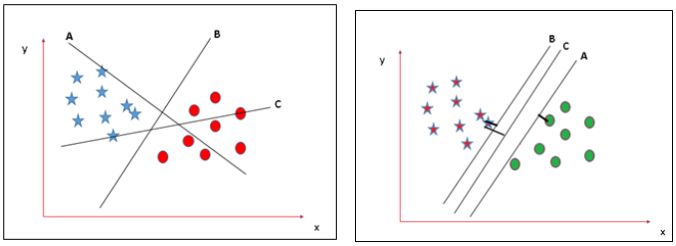

Figura 15. Gráfica ejemplo del modelo de Maquina de soporte vectorial [4].

4.4.3 Redes neuronales

Es una analogía matemática al sistema nervioso humano. Una red neuronal está formada por unidades de procesamiento de información interconectadas (Figura 16).

Las reglas de aprendizaje se pueden utilizar junto con el método de error de propagación inversa. La regla de aprendizaje se utiliza para calcular el error en la unidad de salida. Este error se propaga hacia atrás a todas las unidades de manera que el error en cada unidad es proporcional a la contribución de esa unidad al error total en la unidad de salida. Los errores en cada unidad se utilizan para optimizar el peso en cada conexión. [6]

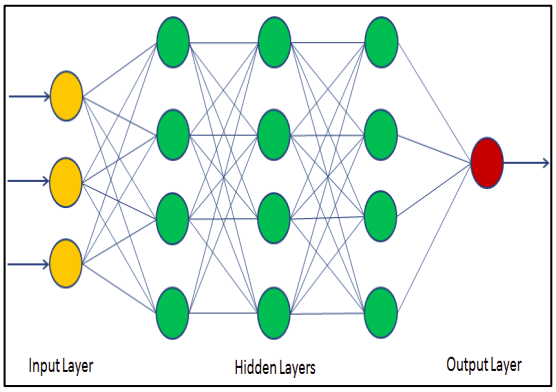

Figura 16. Estructura de las redes neuronales

3. Implementación

Para la implementación de los algoritmos de Machine Learning se ha utilizado el lenguaje de programación R con el apoyo del software RStudio que es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Además, es gratuito y de código abierto. [7]

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos.

En este software se realzó la importación de los datos que hemos extraído de todas las variables, se normalizó las variables y se realizó la preparación de la base de datos de entrenamiento y prueba (training 80% – test 20%), luego se aplicaron los algoritmos de machine Learning para que aprenda de los datos de entrenamiento para que finamente con un modelo entrenado poder predecir la variable de susceptibilidad de deslizamientos. (Figura 17)

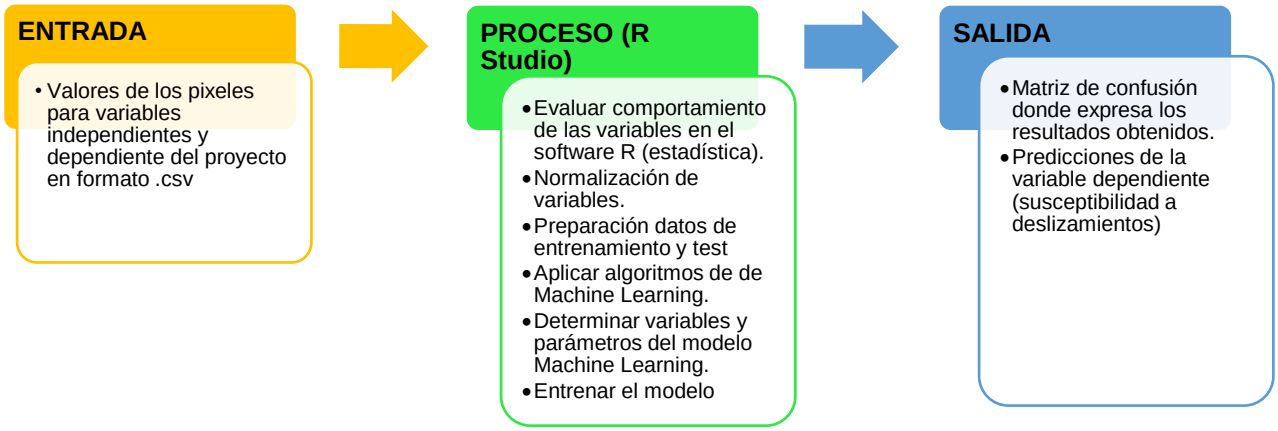

Figura 17. Diagrama de entrada, proceso y salida de datos en el software R Studio.

3.1 Principales instrucciones para la implementación de los algoritmos

Para la implementación de los algoritmos de aprendizaje automatizado se requiere de varias instrucciones las cuales.

Librería/parámetro	Descripción
Regresión logística	
glm	Función para la regresión logística la cual ya está incluida dentro de R estudio
family	Tipo de clasificación en regresión lineal, las más utilizadas son “binomial” y “polinomial”
SVM	
svm	Función de SVM la cual está dentro de la librería (e1071)
kernel	Núcleo utilizado en el entrenamiento y predicción
costo	Es la constante 'C' del término de regularización en la formulación de Lagrange
Redes neuronales	
neuralnet	Librería que contiene las funciones de las redes neuronales.
hidden	Las capas escondidas dentro de las redes neuronales,
act.fct	La función de activación define la salida de una neurona en términos de un campo inducido local.

Tabla 3. Tabla resumen de instrucciones para la implementación de los algoritmos de *Machine Learning.*

4. Resultados y Discusión

4.1 Regresión Logística

Para determinar la relación entre las variables independientes (parámetros morfométricos, vegetación y litología) y dependiente (susceptibilidad de deslizamiento) se elaboró una matriz de correlación entre cada uno de los parámetros (Figura 18) con el propósito de determinar coeficientes que varían desde -0.86 hasta 1. Mientras los coeficientes sean cercanos a 1, el modelo de aprendizaje automatizado tomó mayor importancia a dicho parámetro. Para el caso de análisis, se requiere evaluar la fila “frm” (fenómenos de remoción en masa) y determinar los coeficientes cercanos a 1, obteniendo valores de 0.64 para pendiente, 0.41 para humedad y 0.49 para L2 (areniscas).

Se interpreta que dichas variables tienen mayor relevancia al momento de elaborar los modelos de aprendizaje automatizado (regresión logística, máquina de vectores de soporte y redes neuronales). Por otro lado, los valores cercanos a cero indica que estos parámetros fueron de menos importancia para el modelo, por ejemplo, el plano de curvatura, perfil de curvatura y TWI.

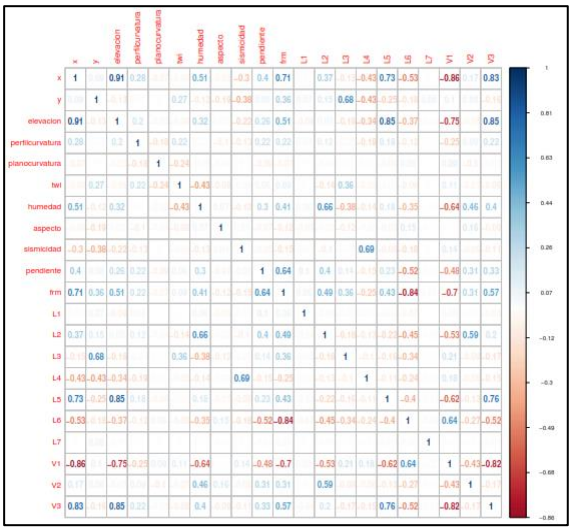

Figura 18. Matriz de correlación de variables.

Análisis de coeficientes y el valor de P(z)

Dentro del modelo de regresión logística una de las maneras para determinar el peso dentro de las variables es a través de los coeficientes y su valor p(z). Valores mayores a 1 indican una correlación de las variables independientes con la dependiente. Mientras que valores menores en p(z) muestran que tiene una baja probabilidad de que suceda de manera aleatoria. La Tabla 4 muestra los coeficientes mayores o iguales a cero y valores p(z) cercanos a cero.

Variable	Coeficiente	Error estándar	P(z)
Pendiente	6.47	1.966	0.00096
Humedad	1.68	1.91	0.37
Elevación	-1.095	5.88	0.98
L2(areniscas)	-7.05	2.22	0.99

Tabla 4. Análisis de los coeficientes en el modelo de regresión logística

En la Tabla 4, se representan en orden descendente los coeficientes que mayor importancia utiliza el modelo para su predicción. La variable pendiente presenta un coeficiente mayor a 1 y un valor de P(z) cercano a cero, por lo que este parámetro sería el ideal para el análisis de susceptibilidad de deslizamientos. Sin embargo, se pueden considerar otros parámetros que respuesta similar a pendiente. La humedad también es una variable a considerar, ya que posee un coeficiente positivo. Mientras que las variables que podrían tomar

menos importancia son elevación y areniscas, por sus parámetros estadísticos. Para la gráfica del modelo de regresión logística se utilizaron pendiente y humedad, con el propósito de evaluar el ajuste de la curva en el modelo propuesto.

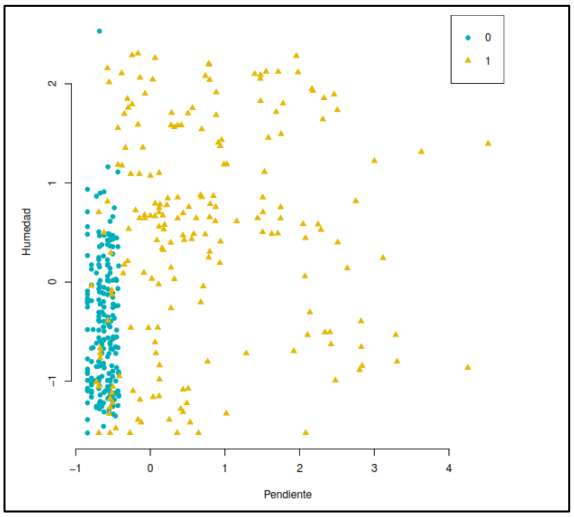

Figura 19. Gráfica del modelo de regresión logística en el eje de las abscisas la pendiente y en el eje de las ordenadas la humedad.

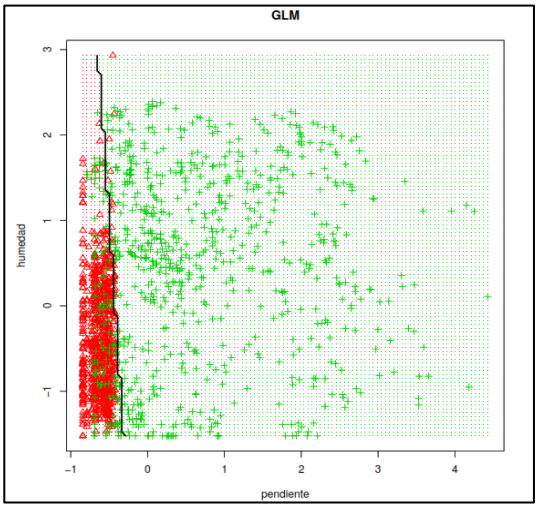

Figura 20. Gráfica GLM del modelo de regresión logística (Pendiente vs. Humedad).

En la figura 20 se observa que el aprendizaje en el modelo de regresión logística presenta una alta precisión ya que al momento de intentar clasificar la pendiente (cruz) y la humedad (rectángulos) traza una línea ideal por la cual separa estos dos parámetros. A partir de la matriz de confusión dentro de los datos de test se pueden determinar los siguientes parámetros estadísticos.

PREDICCIÓN	0	1
0	214	3
1	0	197

Tabla 5. Matriz de confusión regresión logística. (0) No deslizamiento; (1) Deslizamiento.

Exactitud (Accuracy)	**95% CI**	**Sensibilidad**	**Especificidad**
0.9928	(0.97,0.9985)	1,00	0,9850

Tabla 6. Tabla de resultados de predicción regresión logística.

Dentro de la matriz de confusión se obtiene una exactitud de 0.9928, esto quiere decir que apenas en tres valores, el modelo generó un error, es decir realizó una predicción errónea de los resultados. La sensibilidad se define como la relación entre los verdaderos positivos y los verdaderos positivos más falsos negativos. Es la tasa de aciertos cuando la prueba es positiva es decir cuando existe la probabilidad de que ocurra un deslizamiento. Para el modelo de regresión logística tiene un valor de 1. La especificidad es la relación de los verdaderos negativos entre los verdaderos negativos más falsos positivos. Nos indica la capacidad del modelo para dar con los casos negativos, es decir pixeles donde no ocurrieron los deslizamientos.

4.2 Modelo SVM

Para la máquina de vectores de soporte también se utiliza el peso de los coeficientes con el propósito de determinar las variables que tienen mayor ponderación dentro del modelo (Tabla 7). Para visualizar los coeficientes se utiliza la función coef y se obtuvieron los siguientes resultados.

Variable	**Coeficiente**
Pendiente	0.07
Humedad	0.043
Elevación	0.089

Tabla 7. Coeficientes obtenidos con el modelo SVM.

De forma similar al modelo de regresión logística, las variables con mayor ponderación son pendiente, humedad y elevación. Sin embargo, estos valores son negativos por lo que el modelo debe tener una exactitud menor al lineal. Para el análisis la interpretación de las gráficas en el eje de las abscisas se consideró la pendiente mientras que en el eje de las ordenas la humedad.

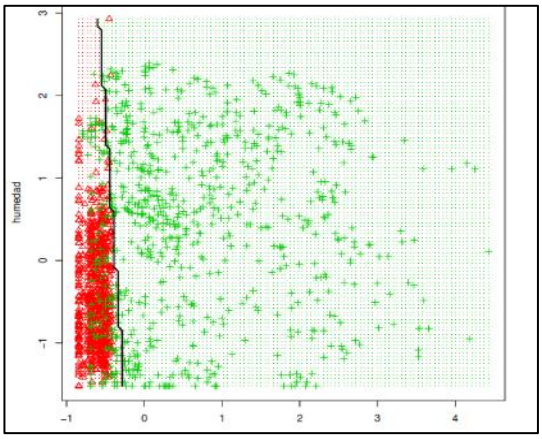

Figura 21. Gráfica del modelo máquina de vectores de soporte utilizando el modelo lineal.

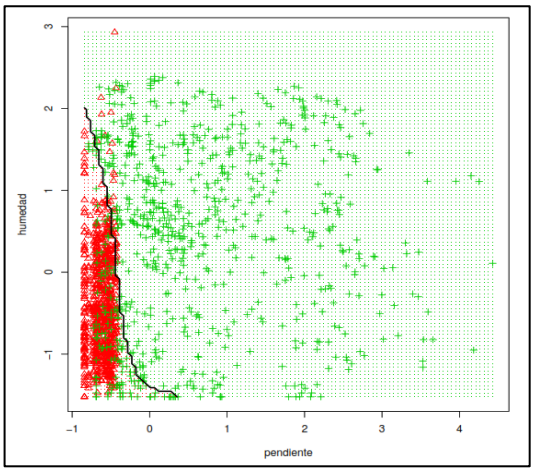

Figura 22. Gráfica del modelo máquina de vectores de soporte utilizando el modelo polinomial.

Dentro de las gráficas de máquina de vectores de soporte de modo experimental se realizó un modelo lineal y polinomial dentro de los cuales se observan significativas diferencias al momento de graficar los hiperplanos. En la gráfica con fórmula lineal (Figura 21), se consideran todos los datos, mientras que en el modelo polinomial (Figura 22) al ser una fórmula cuadrática no considera todos los puntos de análisis.

PREDICCIÓN	0	1
0	212	2
1	3	198

Tabla 8. Matriz de confusión SVM.

En la matriz de confusión (Tabla 8), similar al modelo anterior también se observa una exactitud de 0.993 lo cual también resulta un modelo confiable para la predicción de deslizamientos con una sensibilidad y especificad cercanas a 1.

Exactitud (Accuracy)	**95% CI**	**Sensibilidad**	**Especificidad**
0.9938	(0.97,0.9985)	1,00	0,9870

Tabla 9. Tabla de resultados de predicción SVM

4.3 Modelo de redes neuronales

Para el modelo de redes neuronales no se utilizan coeficientes, se utilizan las capas escondidas (hidden) que para el caso de análisis equivalen a 2, estas capas contienen funciones matemáticas y están diseñadas para obtener un resultado deseado dentro del análisis del modelo. Estas capas son útiles en algoritmos de probabilidad porque toman un dato de entrada y producen un valor de salida entre 0 y 1.

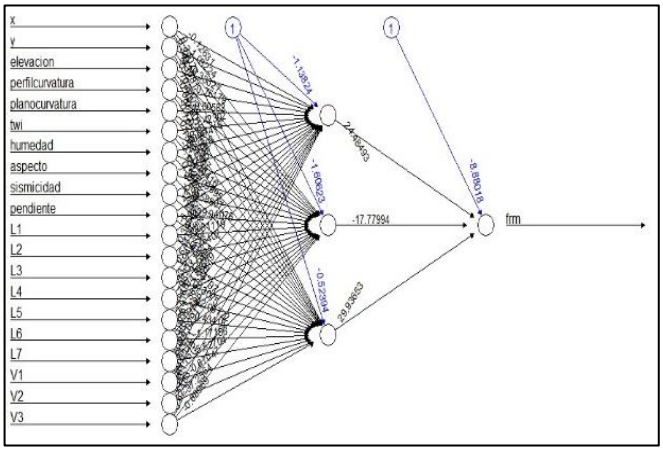

Figura 23. Gráfica del modelo de redes neuronales.

PREDICCIÓN	0	1
0	214	3
1	0	197

Tabla 10. Matriz de confusión redes neuronales.

Dentro del modelo de redes neuronales, se obtuvo una exactitud de 0.992 lo cual resulta un modelo confiable para determinar zonas de posible susceptibilidad de deslizamientos, una sensibilidad y especificidad cercanas a 1 nos indica la capacidad del modelo para predecir casos positivos y negativos de deslizamiento. En la interpretación de la figura 23 se observa que redes neuronales, intenta ajustar la curva para definir una discriminación de las variables, sin embargo, en la parte superior discrimina un solo parámetro obteniendo una inconsistencia en el modelo.

Exactitud (Accuracy)	**95% CI**	**Sensibilidad**	**Especificidad**
0.9928	(0.97,0.9985)	1,00	0,9850

Tabla 11. Tabla de resultados de predicción redes neuronales.

5. Comparación de los tres modelos

Las curvas llamadas ROC (Receiver Operating Characteristic) establecen una relación entre los parámetros de especificidad y sensibilidad de cada uno de los modelos generados (Figura 24), por lo tanto, podemos evaluar el área bajo la curva (AUC) y evaluar cual es el mejor para la aplicación en la predicción de deslizamientos. Para el modelo de regresión logística se obtuvo un valor de 0.990, para la máquina de vectores de soporte 0.988 y para redes neuronales 0.993. Los tres modelos cumplen con una exactitud bastante alta y su diferencia son los decimales. Sin embargo, el que presenta mayor AUC es redes neuronales.

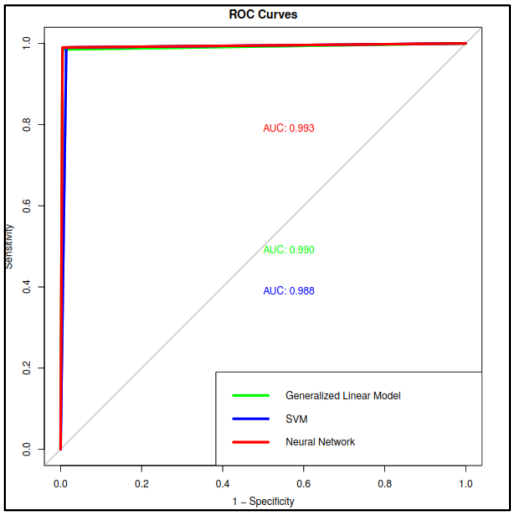

Figura 24. Curvas ROC (Receiver Operating Characteristic)

Una vez obtenidos los resultados para cada uno de los modelos, se los analizó estadísticamente con la finalidad de establecer una clasificación (cinco clases) de la susceptibilidad en intervalos iguales; donde, para los tres modelos se obtuvo valores de 0 y 1 como mínimos y máximos respectivamente (Figuras 25 a 27), para el caso de la media se establece un rango entre 0.4 a 0.6 para los tres casos, mientras que para la moda (curva bimodal para los tres modelos) y mediana se tiene un mismo valor para los métodos de regresión logística y máquina de vectores, lo que no sucede lo mismo para el modelo de redes neuronales dando un valor de 0

Las medidas de forma para, en el caso curtosis se tiene un valor de K<0, dando una distribución de tipo platicúrtica para los tres casos, mientras que para el coeficiente de asimetría se obtiene un valor de CS<0, dando una distribución asimétrica negativa para los modelos de regresión logística (Figura 25) y máquina de vectores (Figura 26), mientras que para el método de redes neuronales (Figura 27) se tiene una distribución asimétrica positiva con CS>0.

ESTADISTICA DESCRIPTIVA DEL MODELO DE REGRESION LINEAL	
Media	0.58301974
Error típico	0.00146469
Mediana	0.99999996
Moda	1
Desviación estándar	0.48779991
Varianza de la muestra	0.23794875
Curtosis	-1.87203281
Coeficiente de asimetría	-0.33664702
Rango	1
Mínimo	0
Máximo	1
Suma	64665.6343
Cuenta	110915
Mayor (1)	1
Menor (1)	0
Nivel de confianza (95.0%)	0.00287078

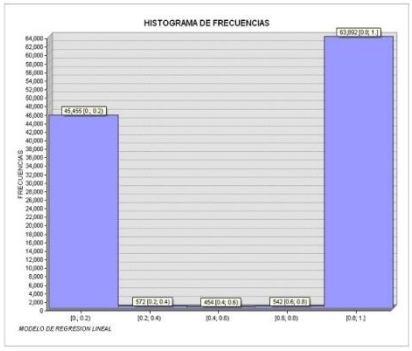

Figura 25. Estadística descriptiva del modelo de regresión logística; histograma dividido para 5 clases de susceptibilidad.

ESTADISTICA DESCRIPTIVA DEL MODELO DEL SVM	
Media	0.62583059
Error típico	0.00145301
Mediana	1
Moda	1
Desviación estándar	0.48390988
Varianza de la muestra	0.23416877
Curtosis	-1.72956116
Coeficiente de asimetría	-0.52006733
Rango	1
Mínimo	0
Máximo	1
Suma	69414
Cuenta	110915
Mayor (1)	1
Menor (1)	0
Nivel de confianza (95.0%)	0.00284788

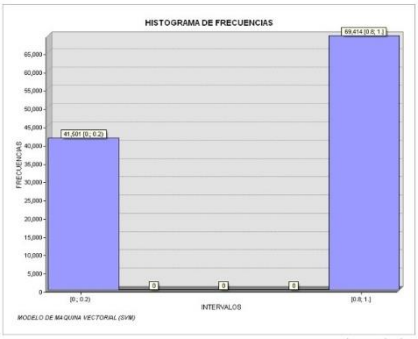

Figura 26. Estadística descriptiva del modelo de SVM; histograma dividido para 5 clases de susceptibilidad.

ESTADISTICA DESCRIPTIVA DEL MODELO DE REDES NEURONALES	
Media	0.46865618
Error típico	0.00149838
Mediana	0
Moda	0
Desviación estándar	0.49901885
Varianza de la muestra	0.24901981
Curtosis	-1.98425438
Coeficiente de asimetría	0.12562406
Rango	1
Mínimo	0
Máximo	1
Suma	51981
Cuenta	110915
Mayor (1)	1
Menor (1)	0
Nivel de confianza (95.0%)	0.0029368

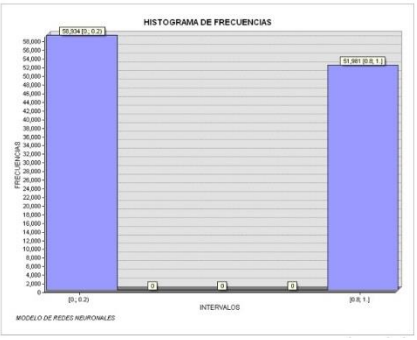

Figura 27. Estadística descriptiva del modelo de redes neuronales; histograma dividido para 5 clases de susceptibilidad.

Las clases (5) obtenidas a partir de distribuciones iguales, otorgan el grado de susceptibilidad para cada mapa (Figuras 28 a 30), siendo los mejores resultados obtenidos de los modelos de regresión logística y maquina vectorial (Figuras 28 y 29) donde se predice una frecuencia cercana tanto en la susceptibilidad de tipo muy baja (0 – 0.2) como muy alta (0.8 – 1), estas frecuencias tienen un nivel de correlación, ya que también estos modelos se correlacionan tanto en medidas de tendencia y de forma (Figuras 28 y 29); pero el mejor modelo que distribuye zonas de susceptibilidad es el modelo de regresión logística (Figura 28), donde se obtienen susceptibilidad medias distribuidas a lo largo de toda la zona de estudio.

Para validar de mejor manara uno de los tres modelos se comparó con un mapa de un análisis de susceptibilidad por el Método de Mora Vahrson, en cual involucra en su estudio un estudio de litología, pendientes, humedad, precipitación y aceleración sísmica horizontal, en este mapa se pueden observar de igual forma cinco clases de susceptibilidad, las cuales se distribuyen de mejor manera a lo largo de toda la zona correlacionando cada uno de los deslizamientos presentes en la zona con una alta susceptibilidad (Figura 31). Este mapa se correlaciona de mejor manera para el mapa obtenido por el modelo de regresión logística (Figura 28), el cual también se correlacionan las zonas susceptibles (susceptibilidad muy alta) con el inventario de deslizamiento y mostrando además nuevas zonas susceptibles a deslizamientos.

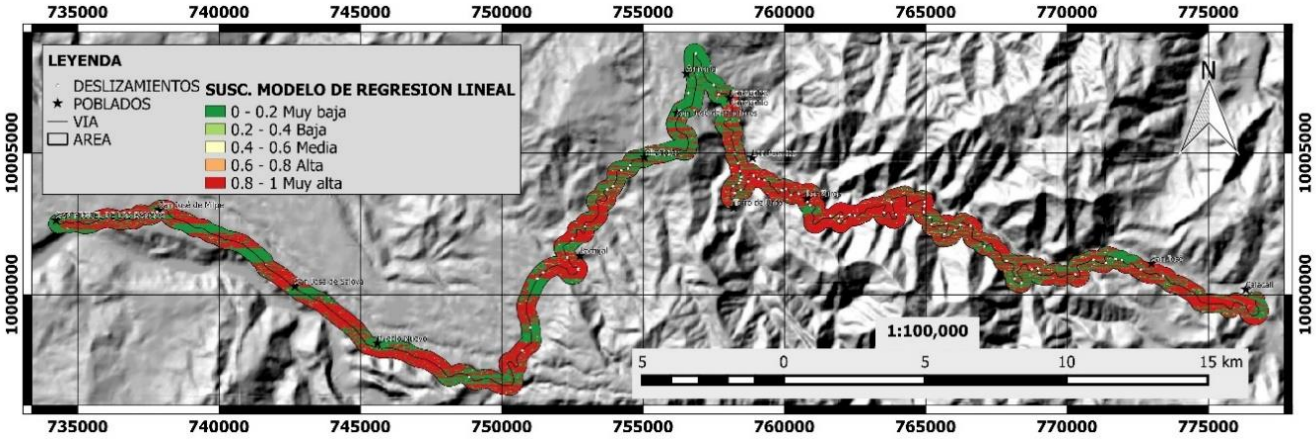

Figura 28. Mapa de Susceptibilidad a Fenómenos de Remoción en Masa en función del Modelo de Regresión Logística.

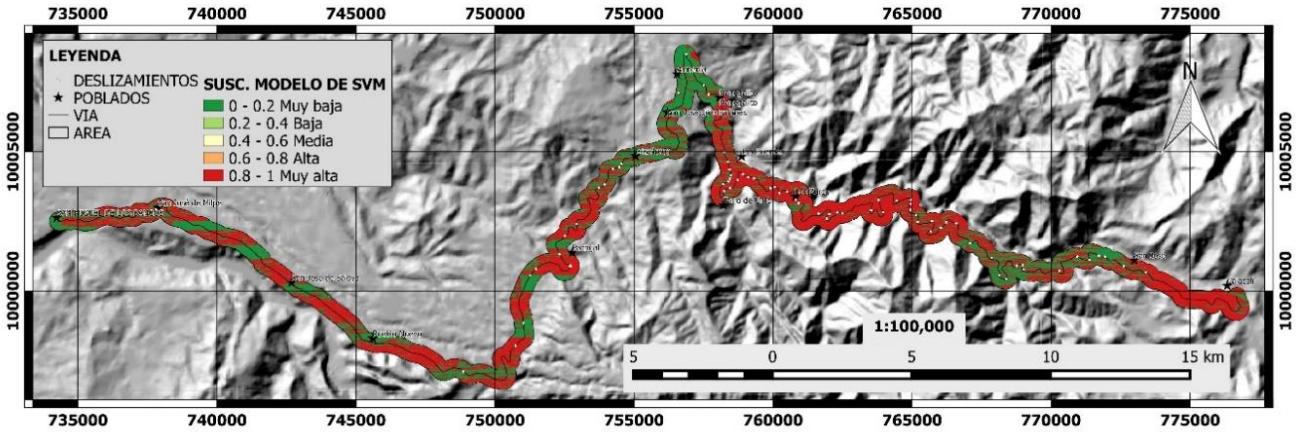

Figura 29. Mapa de Susceptibilidad a Fenómenos de Remoción en Masa en función del Modelo de SVM.

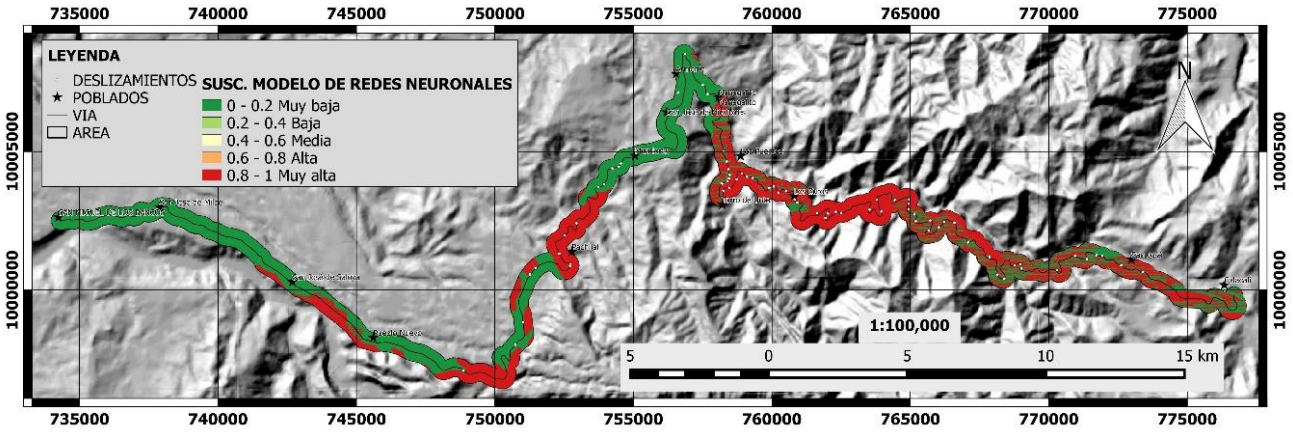

Figura 30. Mapa de Susceptibilidad a Fenómenos de Remoción en Masa en función del Modelo de Redes Neuronales.

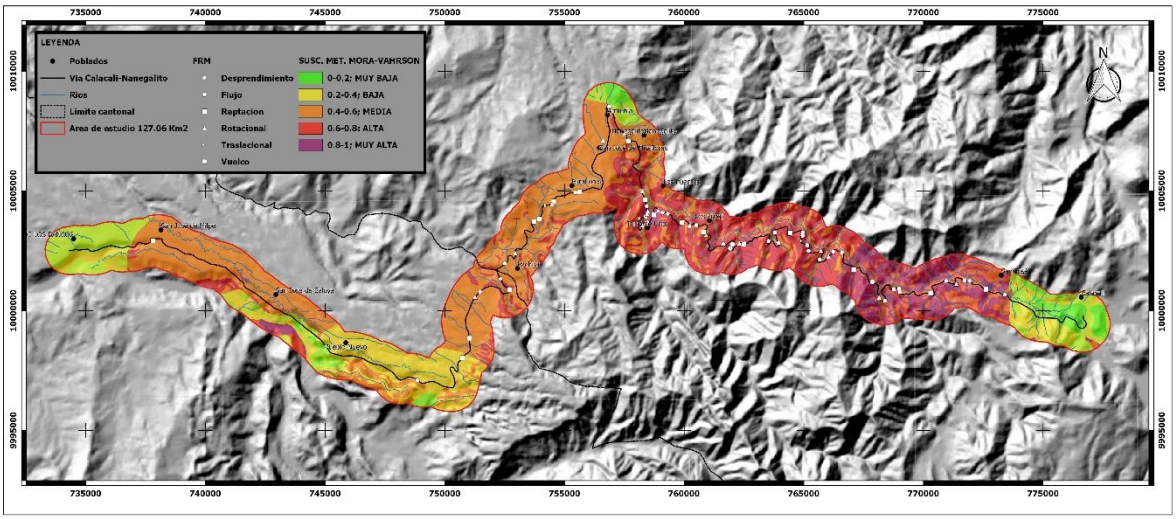

Figura 31. Mapa de Susceptibilidad de Fenómenos de Remoción en Masa en la Vía Calacali- Naneganito. Método de Mora-Varhson.

ALGORITMO	DATOS DE ENTRENAMIENTO		DATOS DE PRUEBA		AUC
	0	1	0	1	
Regresión Logística	856	2	214	3	
	0	796	0	197	
Precisión	99.80%		99.28%		99%
Sensibilidad	99.99%		99.99%		
Especificidad	99.75%		98.50%		
Maquina de vectores de soporte	0	1	0	1	
	838	2	211	2	
	18	796	3	198	
Precisión	98.70%		98.79%		
Sensibilidad	97.90%		98.60%		
Especificidad	99.75%		99%		98.80%
Redes neuronales	0	1	0	1	
	856	1	214	3	
	0	797	0	197	
Precisión	99.30%		99.28%		
Sensibilidad	99.90%		99.90%		
Especificidad	99.97%		98.50%		99.30%

Tabla 12. Resumen de resultados para cada algoritmo de Machine Learning.

La tabla 12 es una comparativa de los tres algoritmos aplicados para determinar la susceptibilidad de deslizamientos, para realizar la comparación se utilizó parámetros estadísticos como la precisión, sensibilidad y especificada. Estos dos últimos parámetros identifican la capacidad de un modelo para detectar casos negativos y casos positivos respectivamente. En los tres modelos el porcentaje es superior al 90% por lo que se consideran excelentes para la predicción de fenómenos de remoción en masa.

Sin embargo, el modelo que más se ajusta a la realidad es el de redes neuronales con un porcentaje de 99.90% en sensibilidad y 99.97% es especificidad. Otro parámetro estadístico importante es el área bajo la curva (AUC) el cual se utiliza como comparativa (curvas ROC) para determinar cuál fue el mejor modelo entrenado. De esta manera las redes neuronales obtuvieron un AUC de 99.30% en comparación con el de máquina de vectores de soporte 98.80% y regresión logística 99%

6. CONCLUSIONES Y RECOMENDACIONES

A partir del análisis de correlación de las variables independientes se determinó que la pendiente, humedad y areniscas son consideradas para la creación de los modelos de aprendizaje automático. Sin embargo, la elevación, perfil de curvatura y twi también influyen. Para determinar el peso de las variables se utilizó el valor P(z) en el modelo de regresión logística el cual representa la probabilidad de que haya seleccionado la variable de manera aleatorio. Por lo tanto, valores

cercanos a cero, representan factores importantes en el momento de definir los factores para la ocurrencia de un deslizamiento.

Dentro del análisis del área bajo la curva en las gráficas ROC, el modelo de redes neuronales presenta una exactitud de 0.993 por lo que estadísticamente es el mejor modelo, pero al momento de elaborar los mapas de susceptibilidad una mejor clasificación es a partir del modelo de regresión logística.

En las gráficas elaboradas, el modelo que mejor se ajusta a una predicción de deslizamientos es el modelo de regresión logística, ya que traza una línea para separar datos positivos y negativos.

El mejor modelo que predice la susceptibilidad a fenómenos de remoción en masa es el Modelo de regresión logística, ya que este se correlaciona con las zonas susceptibles y con los deslizamientos inventariados en la zona. Las zonas de mayor susceptibilidad se encuentran relacionadas con pendientes mayores a 28º y con litologías predominantes de aluviales, calizas, areniscas y lutitas.

La comparación del Mapa de Susceptibilidad del Modelo de Regresión logística con el de Mora-Varhson se puede decir, que la estimación por el modelo de regresión logística arroja zonas mayormente susceptibles a deslizamientos, siendo de gran ayuda para un estudio de campo con el fin de reducir y mitigar posibles riesgos a deslizamientos que pueden ocasionar daño a los pobladores junto a la vía, ya que estos también se encuentran en una zona de riesgo.

Para futuras investigaciones se requiere realizar una validación cruzada de los datos, con el propósito de evitar el sobre entrenamiento de los datos. Además, se recomienda utilizar un ráster con un tamaño de pixel menor a 20 metros con el objetivo de comparar los datos obtenidos.

Para más información del tratamiento de los datos y el código desarrollado en R ingresar al siguiente enlace:

https://uceedu-my.sharepoint.com/:f:/g/personal/dsbustos_uce_edu_ec/ErAVqC6qy_hFnV6ChmMUHboBPYAFg7ySKsrLiMjsMbUJlA?e=lRItFf

REFERENCIAS:

[1] F. Andrade, Análisis de procesos geodinámicos, factores geológicos y medidas de mitigación en taludes inestables a lo largo de la vía Calacalí -Nanegalito. Universidad Central del Ecuador, 2011

[2] Y. Achour, H. Reza Pourghasemi, «How do machine learning techniques help in increasing accuracy of landslide susceptibility maps?» Elsevier, 2019.

[3] Prabhakaran S, «Linear Regression, R-stadistics.» 2020 [En línea]. Available: http://r-statistics.co/Linear- Regression.html.

[4] Meyer D, «Support vector machines, RDocumentation» 2020 [En línea] Available: https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/svm

[5] Navlani A, «Neural Network Models in R, Tutorials». 2020 [En línea]. Available: https://www.datacamp.com/community/tutorials/neural-network-models-r

[6] Verma A, «Building A Neural Net from Scratch Using R - Part 1, R views». 2020 [En línea]. Available: https://rviews.rstudio.com/2020/07/20/shallow-neural-net-from-scratch-using-r-part-1/

[7] Colman A, «Interpretation of the AUC. Datascience». 2020 [En línea]. Available: https://datascienceplus.com/interpretation-of-the-auc/

[8] Zevenbergen, L.W. and Thorne, C.R. (1987) «Quantitative analysis of land surface topography. Earth Surface Processes and Landforms». [En línea]. Available: http://www.etst.com/et_surface/userguide/Raster/ETG_RasterCurvature.htm