Evaluación del Aprendizaje Automático para la predicción de inundaciones

Assessment of Machine Learning in Flood Prediction

JonathanGallardo-Chávez
Universidad Central del Ecuador, Quito, Ecuador
jdgallardo@uce.edu.ec
https://orcid.org/0000-0002-1323-4303

CarolinaJumbo-Olaya
Universidad Central del Ecuador, Quito, Ecuador
cijumbo@uce.edu.ec
https://orcid.org/0000-0003-2691-177X

AdrianaMorales-Loor
Universidad Central del Ecuador, Quito, Ecuador
anmorales@uce.edu.ec
https://orcid.org/0000-0003-0355-3383

ChristianMejía-Escobar
Universidad Centraldel Ecuador, Quito, Ecuador
cimejia@uce.edu.ec
https://orcid.org/0000-0001-6715-191X

Cita del artículo:Gallardo-Chávez J., Jumbo-Olaya C., Morales-Loor A. y Mejía-Escobar C. (2020).Assessment of Machine Learning in Flood Prediction.

Resumen

Las inundaciones son fenómenos hidrológicos que forman parte de la dinámica y evolución de cauces, potencialmente destructivos para poblaciones y campos agrícolas asentados en lasproximidadesde ríos, con posibles pérdidas humanas y materiales. La predicción de este tipo de fenómenos es una actividad permanente de investigación, donde el uso de técnicas de Aprendizaje Automático (Machine Learning) es reciente y una alternativa prometedora a los métodos tradicionales. El presente estudio describe la implementación y comparación de tres algoritmos de aprendizaje automático: Máquina de Soporte de Vectores (SVM), Regresión Logística y Árboles de Decisión, para la predicción de zonas susceptibles a inundación considerando las provincias de Esmeraldas y Manabí en la región costera del Ecuador. Los resultados obtenidosevidencian un alto grado de confiabilidad, sensibilidad y especificidad, indicadoresdecisivospara evaluar labondad de un modelo de aprendizaje automático.Elproducto final consiste en mapas de susceptibilidad a inundaciones elaborados con el soporte de un Sistema de Información Geográfica (GIS) a partir de los algoritmos mencionados, mostrando que las zonas de mayor susceptibilidad están ubicadas hacia el oeste de las provincias de Esmeraldas y Manabí, y que los factores con mayor influencia tienen que ver con la pendiente (menora 5°) y con la distancia a los ríos (menor a los 100 metros).Nuestro trabajo es una aplicación práctica que pretende demostrar la efectividad del aprendizaje automático en un caso real de riesgos naturales, y de esta manera incentivar su aprovechamiento en problemas dentro de este ámbito.

Palabras clave

Aprendizaje automático, educación, inundaciones, predicción, mapa de susceptibilidad.

Abstract

Floods are hydrological phenomena that are part of the dynamics and evolution of streams, potentially destructive for populations and agricultural fields settled in the vicinity of rivers, with possible human and material losses. The prediction of this type of phenomena is a continuous research activity, where the use of Machine Learning techniques is recent and a promising alternative to traditional methods. The present study describes the implementation and comparison of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression and Decision Trees, for the prediction of flood susceptible areas considering the provinces of Esmeraldas and Manabí in the Coast region of Ecuador. The results obtained show a high degree of confiability, sensitivity and specificity, decisive indicators to evaluate the goodness of a machine learning model. The final product consists of flood susceptibility maps elaborated with the support of a Geographic Information System (GIS) from the mentioned algorithms, showing that the most susceptible areas are located towards the west of the provinces of Esmeraldas and Manabí, and that the most influential factors have to do with the slope (less than 5°) and the distance to the rivers (less than 100 meters).Our work is a practical application that aims to demonstrate the effectiveness of machine learning in a real case of natural hazards, and in this way encourage its use in problems within this area.

Keywords

Machine learning, education, floods, prediction, susceptibility map.

1. Introducción

El Aprendizaje Automático (Machine Learning) está presente cada vez más en las actividades cotidianas.Muchas veces sin notarlo aprovechamos sus beneficios, por ejemplo, en la traducción automática de idiomas, las recomendaciones que recibimos de un producto, o la sugerencia de una película. Las aplicaciones crecen en número y abarcan un amplio rango de áreas, incluyendo las denominadas Ciencias de la Tierra y todo lo que ocurre en la naturaleza.

Los fenómenos naturales son procesos que se originan debido a la dinámica externa e interna de nuestro planeta. Entre estos, las inundaciones son uno de los más peligrosos como consecuencia de fuertes lluvias, subidas del nivel del mar, rotura de presas y diques, entre otros factores. En el caso del Ecuador, las provincias del régimen litoral han sufrido constantes pérdidas debido a los procesos mencionados anteriormente, sobre todo en épocas invernales,y cuyas principales afectaciones están relacionadas con la salud, vivienda, agricultura, educación, bienes y servicios.

Por tal motivo, el estudio de este tipo de fenómenos representa un papel fundamental dentro de los campos de la economía, agronomía, geología y gestión de riesgos. Por lo cual, se propone una nueva metodología para análisis de zonas susceptibles a inundación a partir de la implementación de algoritmos de aprendizaje automático, los cuales han ido ganando mayor aceptación en los últimos años como alternativa para el tratamiento de problemas en diversos campos de estudio. La presente investigacióntoma como caso de estudio, la predicción de zonas susceptibles a inundación en las provincias de Esmeraldas y Manabí, ubicadas al NW del Ecuador (Figura1).

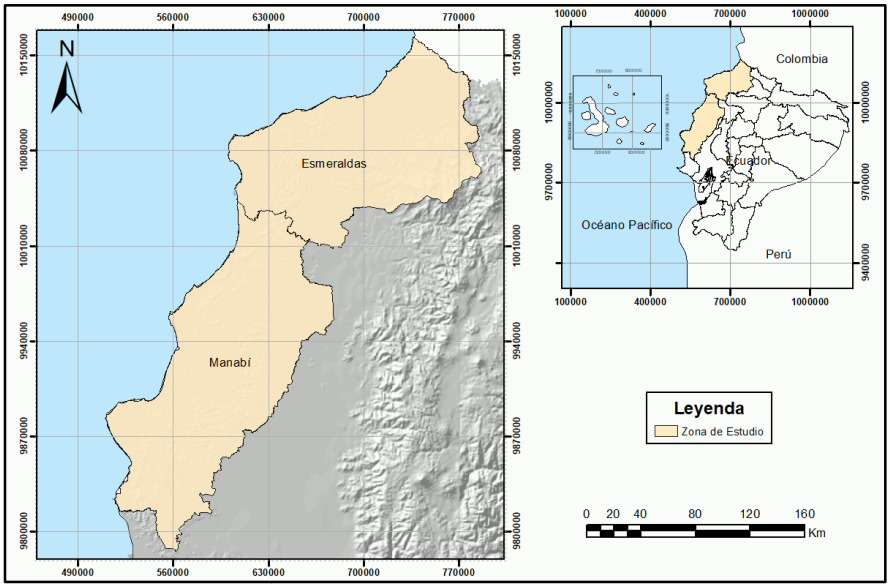


Figura 1. Mapa de ubicación de la zona de estudio.

2. Metodología

La presente investigación sigue un flujograma de trabajo (Figura 2), empezando desde la adquisición de un conjunto de datos extenso y confiable hasta el procesamiento y análisis de los resultados.

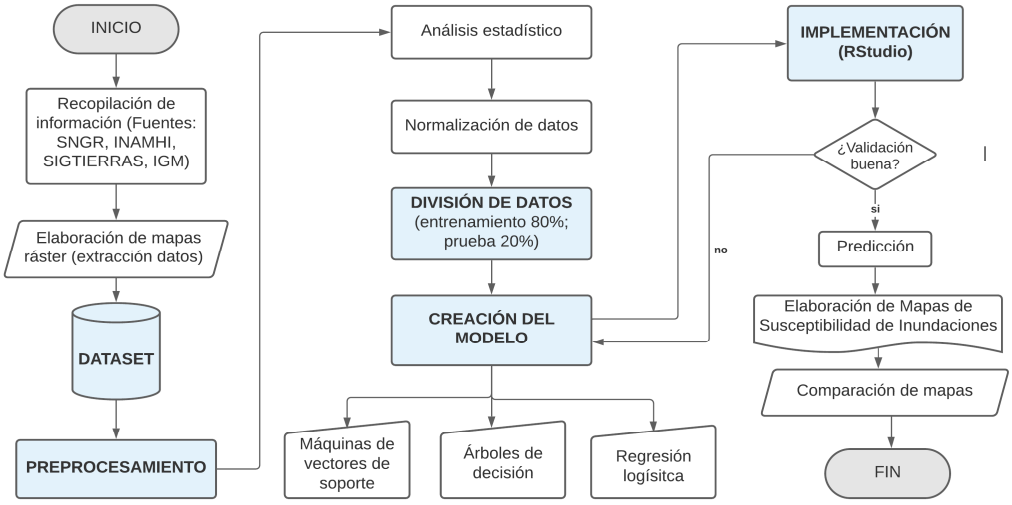


Figura 2. Flujograma de la metodología de trabajo.

2.1. Dataset

La documentación electrónica publicada en sitios de Internet por partede la Secretaría Nacional de Riesgos, el Sistema Nacional de Información de Tierra, y el Instituto Nacional de Meteorología e Hidrología (INAMHI),ha servido como fuentes de información para estructurar y recopilar nuestro conjunto de datos con las variables que se consideran influyentes en un fenómeno de inundación.

Variables

Se estableció la variable dependiente a zonas susceptibles a inundaciones. Y como variables independientes: la ubicación geográfica (coordenadas X, Y), elevación, pendiente, precipitación, distancia al río, uso de suelo y tipo de suelo.Cada unade estas variables es unparámetro esencial de análisis para el estudio de inundaciones (Cuadro 1).

VARIABLE	TIPO	CLASE	ÁREA
Inundabilidad	Cualitativa	Dependiente	Gestión de Riesgo
Coord. X	Cuantitativa	Independiente	Geografía
Coord. Y	Cuantitativa	Independiente	Geografía
Elevación	Cuantitativa	Independiente	Geomorfología
Pendiente	Cuantitativa	Independiente	Geomorfología
Precipitación	Cuantitativa	Independiente	Meteorología
Uso de Suelo	Cualitativa	Independiente	Ambiental
Distancia río	Cuantitativa	Independiente	Hidrología
Litología	Cualitativa	Independiente	Geología

Cuadro 1. Variables consideradas para la zona de estudio.

Las variablesde tipo cualitativo como: zonas inundables, uso de suelo y tipo de suelo deben ser ponderadas, con el fin de estandarizar todos los datos e introducirlos dentro del algoritmo. Para zonas inundables se ha considerado una clasificación binaria de casos positivos y negativos, asignando valores de 1 y 0 para indicar zonas inundadas y no inundadas respectivamente. Seha asignado (1) en base a registros históricos de zonas inundadas, mientras que (0), en base a un análisis estadístico que se detalla posteriormente.La variable de uso de suelo se clasificó en 5 categorías,donde el valor de 1 significa queva a tener menor influencia en la generación de inundaciones, mientras que, el valor de 5 mayor influencia en a inundaciones, tal como menciona (SNGR, 2010). Para la litología, Mora y Varhson (1991), asigna un valor en base a la descripciónlitológica y a las características ingenieriles de las rocas, como la resistencia al corte delmaterial, grado de fisuración, dureza, etc. Por lo tanto, el valor de 1significa que sus propiedades litológicas y mecánicas sean menos susceptibles a inundación, mientras que el valor 5se califica como muy alta susceptibilidad a inundación.

Mapas de variables

Mediante el uso del software de información geográfica QGIS se haelaborado representaciones ráster de las variables dependiente e independientes, con un tamaño de celda de 300m. Se obtuvieron 380600 registros de datos que cubren toda la extensión de la zona de interés, el inventario histórico de inundaciones mostrado en el mapa de la Figura 3-A, permite detectar las zonas más vulnerables a inundaciones en las provincias de Esmeraldas y Manabí. Gran parte de estas zonas de inundaciones son producto del desbordamiento de los ríos, lo que afecta gravemente tanto a las ciudades cercanas como sectores agrícolas.

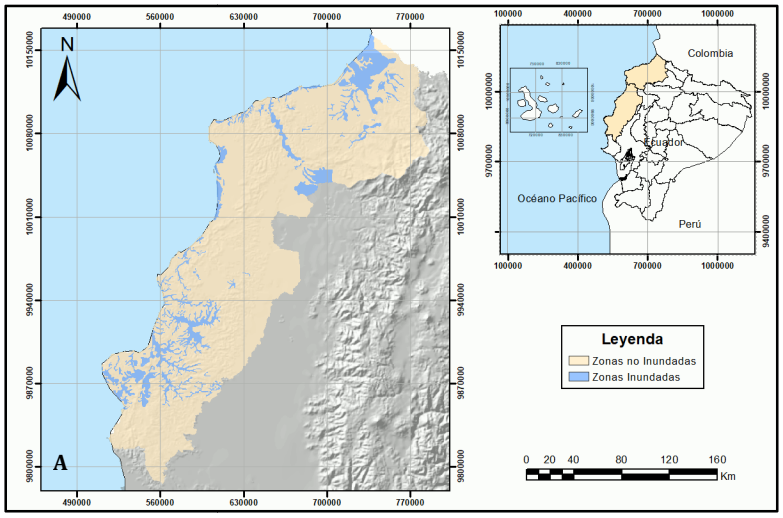


Figura 3. Mapas ráster de variables dependientes e independientes de la zona de estudio. (A) Mapa ráster de zonas inundables en las provincias de Esmeraldas y Manabí.

En el mapa de elevación (Figura 3-B) se identificó el relieve con rangos entre 0 y 3690 m.s.n.m., evidenciando las mayores altitudes hacia el este de la provincia de Esmeraldas y sureste de Manabí. Este modelo digital de elevación fue obtenido de la plataforma Open Topography (https://www.opentopography.org/).

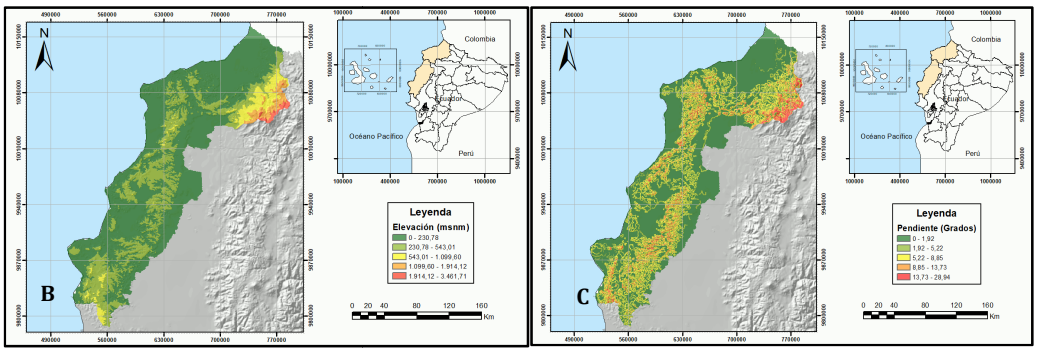
El mapa de pendientes (Figura 3-C) muestra los rangos de inclinación del terreno. Se estableció pendientes mayores a los 5 grados en la parte central de la zona de estudio con tendencia noreste-suroeste, y hacia el este de Esmeraldas. En cambio, los menores valores se localizan principalmente hacia los bordes de las provincias y en zonas de planicies aluviales.

En el mapa de distancia al río (Figura 3-D) se determinó que las zonas más cercanas a los ríos están entre 0 y 176 m de distancia. Sin embargo, hacia el este de la provincia de Esmeraldas existen los mayores valores de distancia, por lo que se asume la ausencia de una red hídrica en este sector.

El mapa de precipitaciones está basado en datos de 29 estaciones meteorológicas durante el períodocomprendido entre los años 1997-2012. Se realizó la interpolación por el método IDW (InverseDistanceWeighting) dando como resultado la Figura 3-E. Con este mapa, se identifica que las mayores precipitacionesse centran hacia el noreste de la provincia de Esmeraldasy centro de la provincia de Manabí.

En el mapa de uso del suelo (Figura 3-F) se observa que la mayor parte del terreno ha sido utilizado para cultivos y pastos plantados, estrechamente asociado a cultivos de café, banano, áreas de camaroneras y sobre todo arrozales.

El mapa litológico (Figura 3-G) determina que para la provincia de Esmeraldas predominan arcillas marinas de estuario además de lutitas y areniscas. En cambio, en la provincia de Manabí prevalecen limolitas, lutitas y areniscas tobáceas. Todas estas rocas han sido ponderadas en base a sus características ingenieriles siendo limolitas, lutitas y areniscas litologías de mayor susceptibilidad a inundación.



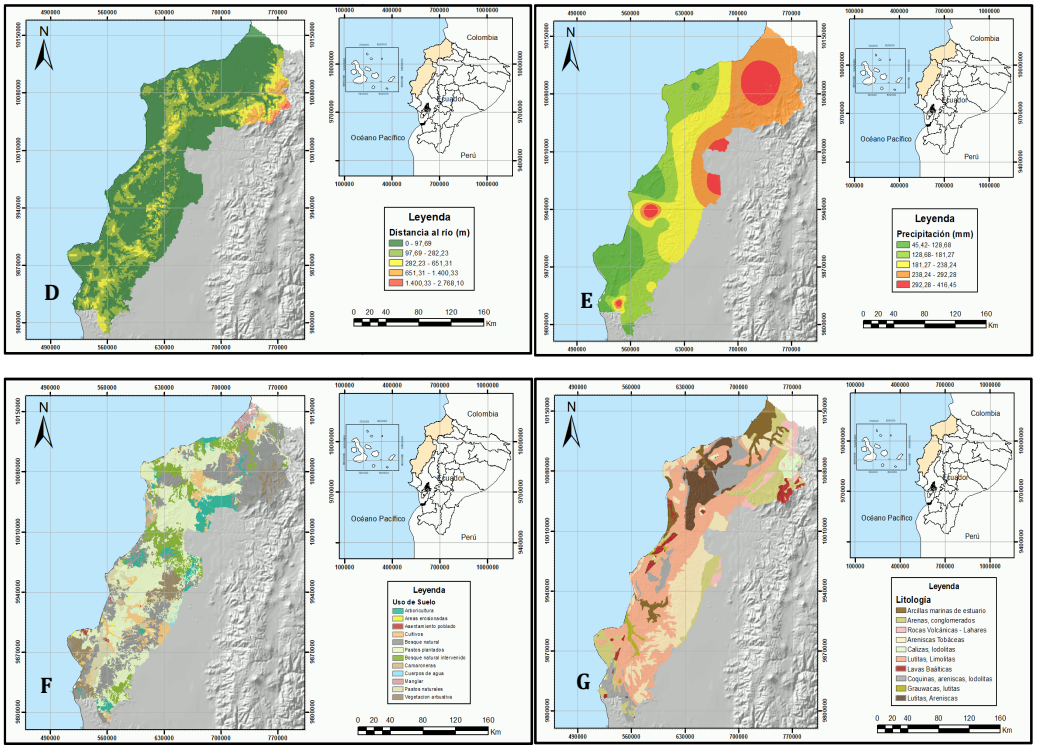


Figura 3. Mapas ráster de variables dependientes e independientes de la zona de estudio. (B) Mapa de Elevación. (C) Mapa de Pendientes. (D) Mapa de Distancia al rio. (E) Mapa de Precipitaciones. (F) Mapa de Uso del suelo. (G) Mapa litológico.

2.2. Preprocesamiento

Esta sección es trascendentalya que debemos cuidar que los datos que serán utilizados en el proceso deentrenamientosean los adecuados y con el formato correcto.

Análisis estadístico

Para conocer el comportamiento de las variables de estudio,eidentificar la existencia de valores atípicos (outliers), realizamos un análisis exploratorio de datos que se resume en el Cuadro 2.

Variable	Mínimo	Máximo	Mediana	Media	Desviación Estándar	Valores atípicos
Coordenada X (m)	510414	786114	634314	642091	66761.8	-
Coordenada Y (m)	9785668	10161868	9999568	9991017	93914.82	-
Elevación (m)	0	3259.30	198.50	258.70	295.88	1000-3259.30
Pendiente (°)	0	28.94	1.31	3.17	3.88	22-28.94
Distancia río (m)	0	3695.10	181.74	263.58	350.52	1100-3695.10
Precipitación(mm)	47.87	417.91	196.10	205.49	67.88	-

Cuadro 2. Indicadores estadísticos de las variables de estudio

En laevaluación de la dispersión de datos se identifican tres variables heterogéneas con valores atípicos como son la elevación, pendiente y distancia al río.El resto de las variables se consideran homogéneas.

Los valores atípicos uoutliers se identificaron en base a los diagramas de cajas y bigotes mediante la estadística obtenida en el software RStudio. Estos valores no fueron omitidos porque pueden causar desviaciones al momento de ejecutar los algoritmos de machine learning.

Normalización de datos

Las diferentes escalas de valores que manejan las diversas variables, así como sus unidades de medida específicas, pueden incidir en predicciones equívocas. Por lo tanto, es importante llevar a cabo un proceso de estandarización, también denominado normalización.

Son varios los métodos que se pueden aplicar. Tal como el escalado estándar, donde a cada dato se le resta la media y se divide por la desviación estándar; o escalado a partir de los valores máximos y mínimos de cada variable. Hemos aplicado este último método de normalización,obteniendo como resultado un conjunto de datos con valores entre0 y 1, listos para usar dentro de la implementación.

2.3. División de Datos

Una vez normalizados los datos, se debe subdividir al conjunto resultante en dos: uno de entrenamiento (training) y otro para prueba (test). Para el presente caso de estudio, se dividió en una proporción de 80-20%, respectivamente, de acuerdo con las prácticas recomendadas dentro del aprendizaje automático.

El conjunto de entrenamiento sirve a los algoritmos para aprender de los datos, ya que proporcionamos la respuesta y el algoritmo intenta llegar a esa respuesta ajustando parámetros y coeficientes. Una vez que se determinan los parámetros y coeficientes más adecuados, se utiliza el modelo para generar la respuesta con los datos del conjunto de prueba, contrastar con las respuestas conocidas y así evaluar un nivel de precisión del modelo.

El conjunto de datos debe satisfacer dos requisitos: ser extenso y confiable. Nuestro datasetfue elaborado a partir de un registro histórico sobre eventos de inundación en la costa ecuatoriana.De esto se obtuvo un total de 42000 datos positivos.En cambio, los datos negativos o aquellos en los que se estableceque no va a ocurrir el fenómeno, fueron extraídos considerando las variables de pendiente mayor al 6% y distancia al río mayor a 200 m.Estos criterios asignan un caso negativo, ya que bibliográficamente la mayoría de las inundaciones que se han registrado, están asociadas a zonas planas o relativamente planas con pendientes entre 0-4%, y desbordamientos de ríos que alcanza un radio de hasta 150 m.

Para un correcto funcionamiento de los algoritmos,la distribución de registros positivos y negativos debe ser lo más equilibrada posible, evitando sesgos a favor de un caso u otro.En consecuencia, se tomaron 35051registros de zonas no inundados (0) y 34990 registros de zonas inundadas (1).Con esto, existe una diferencia muy poco significativa entre ambos casos, irrelevante en la práctica.

2.4. Creación del modelo

El Aprendizaje Automático es un tipo de Inteligencia Artificial que facilita a las computadoras la capacidad de aprender, sin ser programadas explícitamente (Valdez, 2018). Se han seleccionado 3 algoritmos de predicción, cada uno con suscaracterísticas y funciones específicas, mismos que se describen a continuación:

Máquinas de Vectores de Soporte (SVM)

Este algoritmo pertenece a una clase de aprendizaje supervisado, en donde a un conjunto de datos se les realiza entrenamientos para etiquetar las clases.Intuitivamente este algoritmo representa los puntos de la muestra en el espacio y separa las clases en 2 espacios lo más distantes posibles mediante un hiperplano de separación, tal como menciona Amat (2017).La instrucción en lenguaje R para su implementación es:

$$library(e1071)$$

$$modelo\_SVM \ <- svm(\ Inundabilidad \sim ., data = training\_set, kernel =' \ radial', gamma = 1, cost = 10)$$

Donde:

- *Librería (e1071)*: paquete que contiene el algoritmo de SVM.
- *svm()*:instancia del algoritmo SVM para un problema de clasificación.
- *kernel*: establece una separación lineal o radial
- *~.* :comando para utilizar todas las variables independientes.
- *gamma*: distancia entre las observaciones que separan los subespacios del SVM
- *cost*: es el peso que le damos a cada observación a la hora de clasificar.

Nota: *cost* y *gamma* son los parámetros que deciden el desempeño de un modelo SVM. Su funcionamiento se detalla en la sección de discusión.

Regresión Logística

Es un método que ha sido utilizado ampliamente en los problemas de clasificación. Amat (2016), explica que el fundamento se basa en la relación entre la variable dependiente, que se desea predecir, y una o más variables independientes a partir de una función sigmoide para devolver un valor de probabilidad que posteriormente puede ser asignado a dos o más clases. El algoritmo utilizado es:

$$modelo\_RL \ <- \ glm\ (Inundabilidad \sim ., data = training\_set, family = 'binomial')$$

Donde:

- *glm():*función utilizada para un modelo de regresión logística.
- *data*: conjunto de datos de entrada (entrenamiento)
- *family*: número de clases, en este caso, clasificación binaria
- *~.* :comando para utilizar todas las variables independientes.

Nota: Los parámetros aplicados en la fórmula para el caso de inundaciones se detallan en la discusión de resultados.

Árboles de Decisión

Bouza, et al. (2016) menciona que es un modelo que predice la variable dependiente con base en el aprendizaje de reglas de decisión inferidas desde las características que poseen los datos. Los árboles de decisión son llamados así porque su representación incluye elementos como: raíz, nodos, ramas y hojas. El nodo raíz y los nodos internos del árbol corresponden a una prueba del valor de las propiedades y las ramas nodo son identificadas mediante los posibles valores de la prueba. En los nodos hoja del árbol se especifica el valor que se debe producir en el caso de alcanzar dicha hoja. El algoritmo utilizado es:

$$library(rpart)$$

$library(rpart.plot)$

$$Modelo\_TREE < -rpart(Inundabilidad \sim ., data = training\_set, method = 'class')$$

Donde*:*

- *library(rpart)*: paquete que incluye árboles de decisión a partir de tablas.
- *library(rpart.plot)*: paquete para graficar los árboles de decisión
- *rpart()*: instancia de un algoritmo de árbol de decisión
- *~.* : comando para utilizar todas las variables independientes.
- *method*: clasificación (class) o regresión (Anova).

Nota: Los hiperparámetros principales para el algoritmo de árboles de decisión son *minsplit*, *minbucket*, parámetro de complejidad (*cp*) y *maxdepth*, los cuales se detallan en la discusión.

3. Implementación

3.1. Lenguaje de programación R

R es un lenguaje de programación de carácter libre diseñado específicamente para el análisis estadístico de datos y quepermite generar prototipos de modelos con excelentes resultados, como son los algoritmos de Machine Learning (Valdez, 2018). Para facilitar la programación, se ha aprovechado el software gratuito RStudioversión 1.3.1073 instalado en computadoras que están al alcance de cualquier usuario, es decir, no se necesita tener la computadora más nueva, o más cara para realizar trabajos de aprendizaje automático.

3.2. Diagrama de procesamiento

La Figura 4 muestra las fases de procesamiento que se realizan al trabajar con el software de RStudio. El principal insumo es el DATASET con las variables que deben analizarse estadísticamente, como paso previo al procesamiento de cada uno de los algoritmos propuestos. Los resultados deben ser evaluados mediante una matriz de aciertos y fallos; y finalmente se prueba los datos seleccionados de test y entrenamiento en dicha matriz. De esta manera se generan los resultados para cada algoritmo y su porcentaje de confiablidad. Por último, se analiza el algoritmo óptimo para estimar las zonas susceptibles a inundaciones en el área de estudio.

Con los coeficientes de predicción obtenidos de cada modelo, mediante el comando *write.csv* se exporta la tabla de resultados a una hoja de Excel. Se incluye las coordenadas para llevar el archivo a un sistema de información geográfica, los cuales son convertidos en ráster de píxel 300 x 300m. Finalmente se clasifica las zonas susceptibles a inundaciones en 5 categorías.

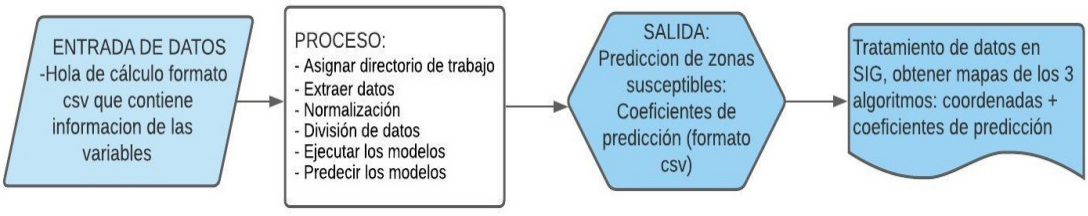


Figura 4. Diagrama de procesamiento para el análisis en RStudio y QGIS

3.3. Disponibilidad del código

La implementación completa de cada uno de los algoritmos utilizados para la predicción de susceptibilidad a inundaciones se encuentra disponible para acceso y descarga en la siguiente carpeta: https://1drv.ms/u/s!AtlWlChdOmtJix67kX9zVYkqOo7O?e=tOBk0x

4. Resultados y Discusión

La ejecución de cada algoritmo trae consigo varios indicadores que permiten evaluar la calidad de las variables y por ende del modelo. Una vez dividido el dataseten uno de entrenamiento y otro de prueba, se ejecutan los algoritmos,para que la computadora aprenda y logre predecir zonas susceptibles a inundaciones. Con la finalidad de comprobar la precisión de los modelos, es fundamental realizar un análisis de la matriz de aciertos y fallospara datos de entrenamiento y de prueba. Además, se examinan los valores de la confiabilidad (precisión), la sensibilidad (casos positivos), y la especificidad (casos negativos), incluidos en el Cuadro 3.

Algoritmo	Datos de Entrenamiento			Datos de Prueba		
Máquina de Vectores de Soporte		**0**	**1**		**0**	**1**
	0	35047	58	**0**	8762	9
	1	4	34932	**1**	1	8738
	Confiabilidad		99.91%	Confiabilidad		99.91%
	Sensibilidad		99.91%	Sensibilidad		99.89%
	Especificidad		99.83%	Especificidad		99.91%
Regresión Logística		**0**	**1**		**0**	**1**
	0	35034	47	**0**	8755	10
	1	17	34943	**1**	2	8739
	Confiabilidad		99.91%	Confiabilidad		99.93%
	Sensibilidad		99.95%	Sensibilidad		99.91%
	Especificidad		99.87%	Especificidad		99.91%
Árboles de decisión		**0**	**1**		**0**	**1**
	0	35050	468	0	8763	98
	1	1	34522	1	0	8649
	Confiabilidad		99.33%	Confiabilidad		99.44%
	Sensibilidad		99.99%	Sensibilidad		99.99%
	Especificidad		98.66%	Especificidad		98.88%

Cuadro 3. Resultados de la ejecución de los algoritmos.

En cuanto al algoritmo SVM, se ingresan todas variables, previamente entrenadas y normalizadas, en donde el algoritmo intuitivamente representa a todos los puntos de muestra en el espacio. Un SVM separa en dos clases mediante un hiperplano de separación, en este caso valores 1 ó 0 en función si corresponde a una zona o no susceptible a inundación respectivamente. Para aquello, se ha seleccionado un costo de 10, un kernel*radial* ya que los datos no son linealmente separables, y se ha adjuntado el parámetro gamma de 0.5.El valor de gamma mientras sea más bajo, la línea de límite será más suave;y si es más grande posiblemente se obtendrá decisiones fluctuantes y sobreajustes en el modelo.Cabe recalcar que el valor de gamma es 1 por defecto. Como

resultado se ha obtenido un SVM de 464 vectores, comprobando que para los datos de entrenamiento el grado de confiabilidad es del 99.91%, y para datos de prueba el grado de confiabilidad es del 99.91%. Por lo que se establece como un modelo muy preciso para la resolución de problemas de clasificación.

Para el algoritmo de regresión logística el peso o grado de importancia de las variables es fundamental para obtener buenos resultados. Se analiza factores como los coeficientes de estimación asignado a cada variable, el error estándar, y la relación entre el coeficiente de regresión y su error, al aplicar la función *summary*. Como resultado de este análisis se logra evidenciar a las variables de elevación, pendiente, distancia al río y precipitación como las de mayor peso o preponderancia para realizar la predicción. Por lo tanto, se procede a ejecutar un nuevo algoritmo utilizando únicamente estas variables de mayor peso con la finalidad de mejorar el rendimiento del algoritmo. De tal manera, para los datos de entrenamiento se obtuvo un grado de confiabilidad del 99.91%, y para los datos de prueba el grado de confiabilidad es del 99.93%. Por lo cual se convierte en un modelo bastante preciso para la predicción de zonas susceptibles a inundaciones. Estos resultados también pueden ser expresados mediante la fórmula general de la regresión logística:

*Inundabilidad = 62.87–25.65*Elevación-127.66*Pendientes-352.87*Distancia al río-8.99*Precipitación*

El algoritmo de árboles de decisión toma en cuenta las variables de mayor importancia para representar el grafico (Figura 5), por ello se destacan las variables de pendiente y distancia al río. Es muy importante notar el uso de los hiperparámetros, debido a que estos ramifican o segmentan la distribución de los árboles de decisión. Los hiperparámetros más importantes son *minsplit*, *minbucket* y el parámetro de complejidad conocido como *cp*.El minsplit nos indica el mínimo de observaciones en un nodo para este se divida (mínimo para que sea padre). El minbucket representa el mínimo de individuos para que sea hijo y el *cp* nos indica la complejidad del árbol y generalmente se expresa en porcentaje. En la investigación se ha colocado el *cp* de -1 para proceder a la poda del árbol, obteniendo un gráfico de mejor distribución y comprensión para el lector. Debido a la gran extensión de datos se debe rescatar que varios nodos no se graficarán luego de dicha poda. Es importante analizar el número de divisiones y los errores mediante el comando *printcp*; con el fin de escoger el valor del *cp* menor. Al evaluar el algoritmo nos indica altos valores de confiabilidad (99.33%) para los datos de entrenamiento y un 99.44% de confiabilidad para los datos de prueba.

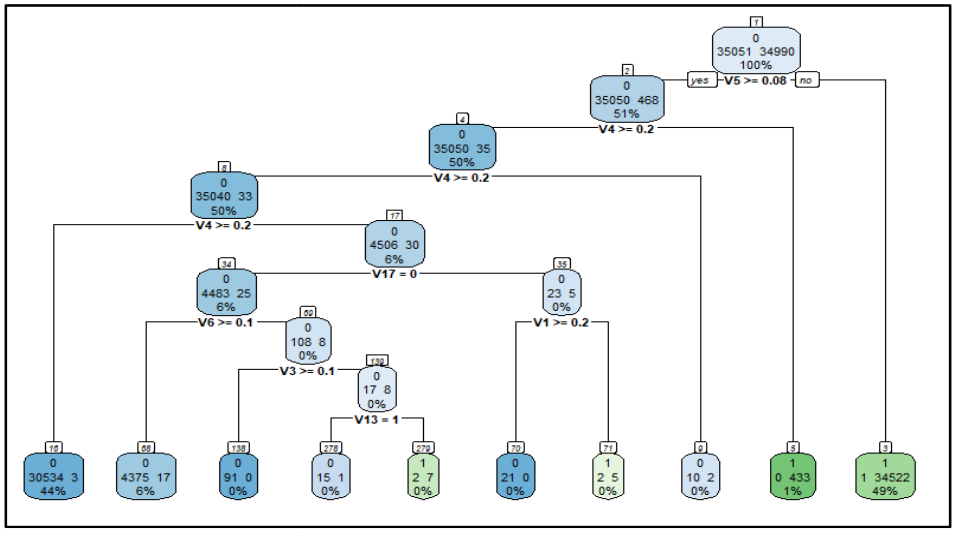


Figura 5. Gráfico del algoritmo árboles de decisión. Tono azul intenso es indicativo de pureza de los datos (zonas no inundadas); tonos verdes asociado a zonas inundadas. La variabilidad en la intensidad del tono está relacionada a la proporción de los datos, es decir el contraste entre ellas.

En resumen, a partir del Cuadro 3 se deduce que los algoritmos tienen una respuesta bastante parecida, siendo los tres buenos para predecir zonas susceptibles a inundaciones. Sin embargo, para un correcto análisis de los resultados de cada uno, y poder comparar entre ellos, una herramienta fundamental son las curvas ROC (Figura 6). Las curvas ROC establecen una relación entre la especificidad y sensibilidad, dando como respuesta un área bajo la curva (AUC). La tendencia o forma de las curvas se torna en forma de "bisagra" debido a que los valores de área bajo la curva son muy elevados, teniendo en cuenta que mientras la sensibilidad es mayor y el valor de *1-especificidad* es menor los resultados serán mejores.En consecuencia, los tres modelos cumplen con una precisión bastante alta, encontrándose en el orden de 0.99. Sin embargo, los árboles de decisión presentan un valor menor en comparación con los otros dos modelos ejecutados (regresión logística y máquina de soporte de vectores).

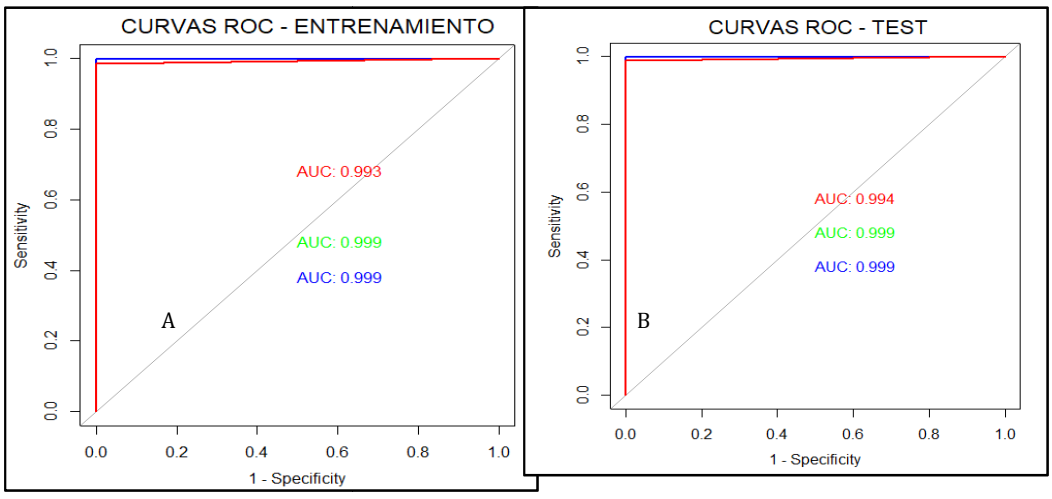


Figura 6. Curvas ROC. 6-A: Datos de Entrenamiento. 6-B: Datos de Prueba. La curva roja representa el modelo de árboles de decisión; la curva azul el modelo de máquina de soporte de vectores; y la curva verde el modelo de regresión logística.

Comprobados los resultados de cada uno de los modelos, se procede a realizar la predicción de zonas susceptibles a inundación mediante la instrucción *predict* que recibe dos argumentos principalmente. Inicialmente el modelo que usaremos para la predicción, el cual ya ha sido anteriormente entrenado y ajustado; y el conjunto de datos de partida de 380600 registros. Los valores esperados se encuentran entre 0 y 1, y para realizar la predicción utilizando el modelo de regresión logística y máquina de soporte de vectores se debe usar el comando *response*. Para el algoritmo árboles de decisión se aplica el comando *prob*. Los resultados de los algoritmos han sido exportados al software de información geográfica QGIS para obtener un modelo ilustrativo y categorizar las zonas de susceptibilidad a inundación.Los mapas obtenidos (Figura 7), han sido clasificados en cinco categorías (muy baja, baja, media, alta y muy alta) utilizando la clasificación estadística *natural breaks*, donde las zonas más susceptibles a inundaciones se encuentran hacia el oeste de las provincias de Esmeraldas y Manabí, los cuales se asocian a regiones donde la elevación y pendiente son relativamente bajas (Elevación < 500 msnm; Pendiente < 5°); y la distancia al río menor a 100 m, lo cual se debe a que uno de los factoresde mayor preponderancia son los desbordamientos de los ríos durante las épocas de invierno, siendo un factor de influencia las precipitaciones que en estas zonas se caracterizan por valores que se encuentran en el orden de 250-400 mm. Además, el uso de suelo en ciertos sectores ocasiona zonas más vulnerables a inundaciones, teniendo sectores de cultivos y manglares en estas zonas, donde la litología predominante son arcillas marinas de estuario, lutitas y areniscas. En cambio, las zonas de susceptibilidad muy baja se encuentran asociados a sectores de alta montaña, así como parte de la cordillera Chongón Colonche ubicada en la zona costera ecuatoriana.

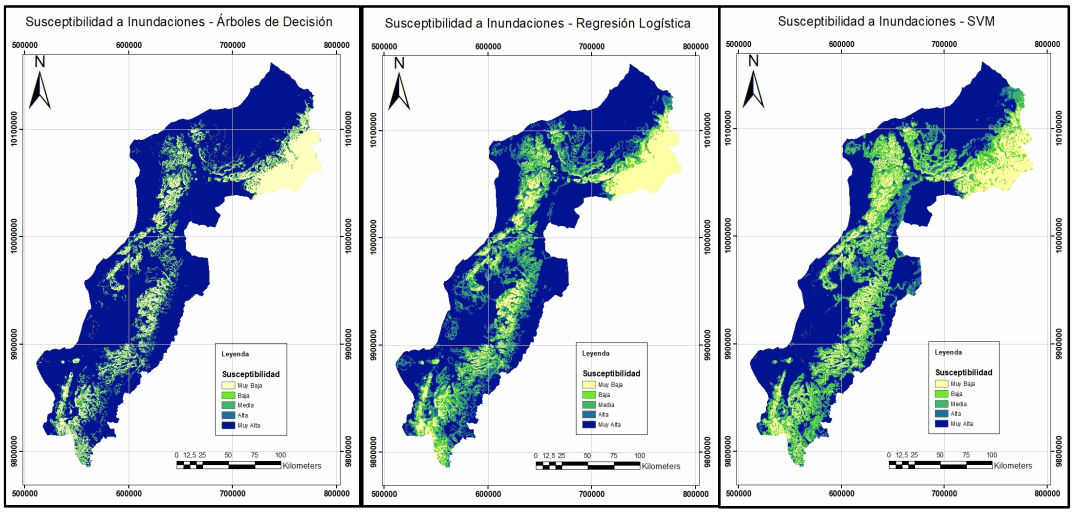


Figura 7. Mapas de Susceptibilidad a Inundaciones de las provincias de Esmeraldas y Manabí. Resultados de los algoritmos de izquierda a derecha: SVM, Regresión Logística y Árboles de Decisión

Finalmente, a partir de los análisis realizados por medio de las matrices de aciertos y fallos, curvas ROC y mapas de susceptibilidad, se logra deducir que los mejores modelos para el presente caso de estudio son la regresión logística y máquina de soporte de vectores, debido a que sus resultados se ajustan a datos bibliográficos sobre inundabilidad (Figura 8) obtenido del mapa de Modelos de los Regímenes de Inundación del Ministerio del Ambiente (2013). En cambio, para el caso de los árboles de decisión, a pesar de que en los resultados se obtuvieron altos grados de confiabilidad (~99%), el mapa de

susceptibilidad resultante muestra mayores zonas de posible inundación, lo cual difiere de los mapas resultantes utilizando los otros modelos.

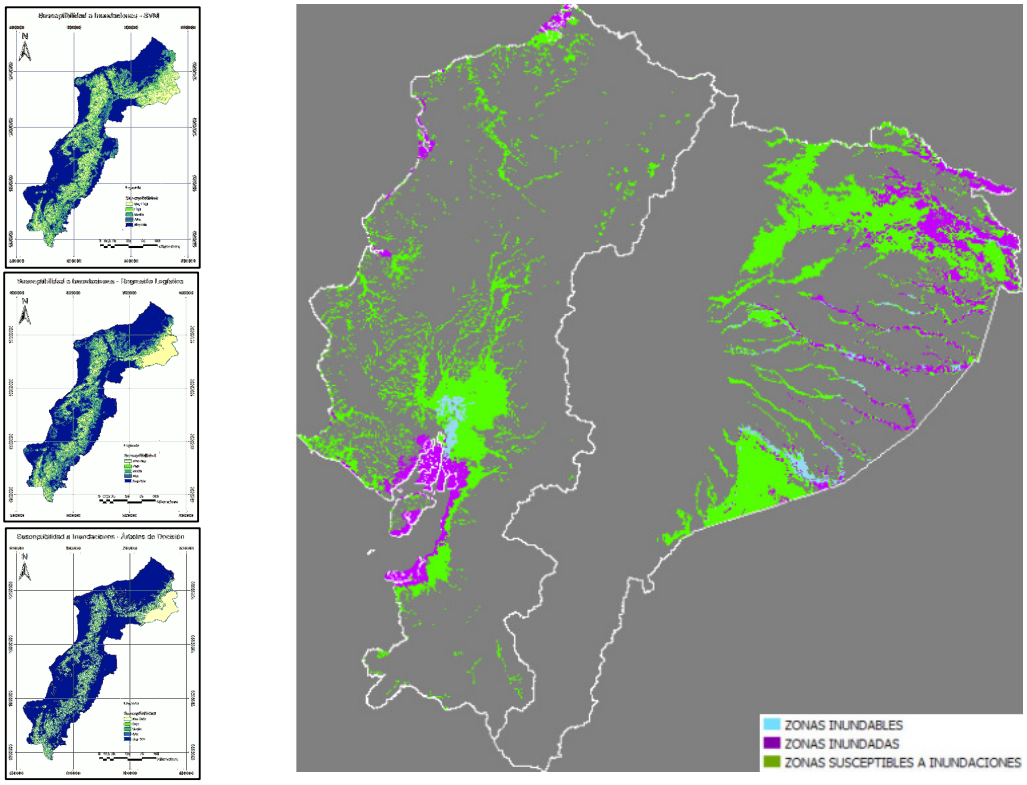


Figura 8. Comparación entre mapas de zonas susceptibles a inundación del Ecuador; mapas obtenidos a partir de los algoritmos de machine learning y Modelos de Regímenes de Inundaciones, Ministerio del Ambiente (2013)

5. Conclusiones y Recomendaciones

La recopilación previa de información para obtener el *dataset* que será entrenado y probado es la clave de éxito, debido a que en ella se exponen las variables que los usuarios consideran relevantes para dicho análisis, por ello se debe ser meticuloso al momento de seleccionarlas, además debemos tomar en cuenta que mientras más extenso y confiable es el dataset, mejores serán los resultados obtenidos. En el presente estudio se obtuvieron 380600 datos extraídos mediante el software QGIS.

Es crucial realizar un análisis estadístico de los datos adquiridos, puesto que, el disponer de dicha información nos permitirá interpretar de una manera lógica la distribución, variabilidad, y agrupamientos de las variables; además nos ayudará a tomar decisiones acertadas para segmentar nuestra base de datos.

LaRegresión Logística, SVM y los Árboles de Decisión, son algoritmos muy efectivos para la resolución de problemas de clasificación, y sin duda para estudios de susceptibilidadainundaciones, en vista al grado de confiabilidad alrededor del 99% como resultado en los tres algoritmos. Sin embargo, al comparar los modelos, a partir de aciertos y fallos, especificidad y sensibilidad, se deduce que el algoritmoóptimo para la predicción de zonas susceptibles a inundación es el de Regresión Logística.

Con los resultados obtenidos en los mapas de susceptibilidad, se identifica que las zonas de mayor riesgo a inundaciones se localizan al oeste de las provincias de Esmeraldas y Manabí, a consecuencia de laspendientesmenores a 5°, elevaciones entre 0 y 300msnm, distancias aríos menores a 100m, y litologías correspondientes a lutitas y areniscas. Ésto coincide en los tres mapas, sin embargo, con respectos a las zonas de menor susceptibilidad a inundaciones, el algoritmo de árboles de decisión predice un menor porcentaje de éstas, lo que se debería a que la precisión, sensibilidad y especificidad es menor en comparación con la regresión logística, y máquina de soporte de vectores.

Los desastres naturales como sismos, erupciones volcánicas y fenómenos de remoción en masa, son eventos geológicos que han sido estudiados con determinación por las grandes pérdidas que han ocasionado, creando diversas metodologías para su evaluación. Mientras que, para eventos de inundaciones pocas han sido las sistemáticas empleadas, en consecuencia, el presente trabajo representa una idea innovadora cuya metodología se ajusta a la predicción de zonas susceptibles a inundaciones; las cuales pueden ser utilizadas para tomar decisiones con respecto a ordenamiento territorial y planes de acción temprana.

Los docentes deben promover nuevas propuestas didácticas de aprendizaje, debido a que la tecnología avanza cada día y con ella su aplicación en los campos de la educación, de manera que, el machine learning o aprendizaje automático representa una herramienta muy favorable para el tratamiento de problemas relacionados con las ciencias de la Tierra; es decir, con estos algoritmos se puede predecir y enfrentar problemas de manera coherente y acertada, gracias a un minucioso análisis de datos por medio de herramientas computacionales como R y RStudio. Además, esta implementación sepuedereplicaren otros campos educativos, con el fin de verificar su validez desde otras perspectivas de enseñanza y aprendizaje.

6. Bibliografía

Amat, R. (2016). Regresión logística simple y múltiple. Ciencias de Datos. cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

Amat, R. (2017). Máquina de Vector Soporte (Support Vector Machines, SVMs). Ciencia de Datos. https://github.com/JoaquinAmatRodrigo/Estadistica-con-R

Bentacourt, G., (2005). Las Máquinasde Soporte Vectorial (SVM's). UTP. ISSN 0122-1701 67. Ingeniería Eléctrica. Universidad Tecnológica de Pereira.

Bouza, C. & Santiago, A. (2014). La minería de datos: arboles de decisión y su aplicación en estudios médicos. modelación matemática de fenómenos del medio ambiente y la salud. isbn: 978-607-7760-60-3. Universidad de Havana

Ministerio del Ambiente (2013). Modelo de los Regímenes de Inundación para la Representación Cartográfica de Ecosistemas del Ecuador Continental. Ecuador. http://app.sni.gob.ec/sni-link/sni/PDOT/NIVEL%20NACIONAL/MAE/ECOSISTEMAS/DOCUMENTOS/Inundabilidad.pdf

Mora, S. &Vahrson, W. (1991). Determinación a priori de la amenaza de deslizamientos sobre grandes áreas, utilizando indicadores morfodinámicos. En: Memoria sobre el Primer Simposio. Bogotá, Colombia. pp. 259-273

Secretaría Nacional de Gestión de Riesgos. Taller II: Unificación de Metodologías para la Valoración de la Amenaza año 2010. Consultado el 25 de junio del 2020https://www.gestionderiesgos.gob.ec/

Valdez, A. (mayo, 2018). Introducción al Machine Learning. Universidad Mayor de San Andrés | UMSA. Consultado el 16 de julio del 2020. https://www.researchgate.net/publication/338518560_INTRODUCCION_AL_MACHINE_LEARNING

7. Presentación de autores

JONATHAN GALLARDO-CHAVEZ. Estudiante de noveno semestre de la carrerade Ingeniería en Geología, de la facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental, de la Universidad Central del Ecuador.

CAROLINA JUMBO-OLAYA. Estudiante de noveno semestre de la carrerade Ingeniería en Geología, de la facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental, de la Universidad Central del Ecuador.

ADRIANA MORALES-LOOR. Estudiante de noveno semestre de la carrerade Ingeniería en Geología, de la facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental, de la Universidad Central del Ecuador.

CHRISTIAN MEJIA-ESCOBAR. Docente de la Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental, de la Universidad Central del Ecuador.