PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Clasificación del territorio peruano de acuerdo con su potencial de agua subterránea utilizando algoritmos de aprendizaje automatizado

Tesis para optar el título profesional de Ingeniero Civil

AUTOR:

César Augusto Portocarrero Rodríguez

ASESOR:

Gustavo Martín Larrea Gallegos

Lima, 25 de noviembre de 2020

Resumen

El agravamiento del estrés hídrico tanto en el sector urbano como en el rural motiva cada vez más a los tomadores de decisión a impulsar la explotación sostenible de este recurso. Para ello, se requiere conocer con certeza los emplazamientos con un mayor potencial de explotación. Para hacer frente a este problema sin recurrir a perforaciones directas, la presente investigación tiene como objetivo principal explorar el potencial hidrológico subterráneo del Perú correspondiente a acuíferos de baja profundidad mediante la aplicación de modelos de clasificación de bosques aleatorios y redes neuronales, dos algoritmos de aprendizaje automatizado. Esta rama de la inteligencia artificial permite generar modelos multidimensionales y con variables complejas sin efectuar presuposiciones estadísticas. Para explicar el potencial de agua subterránea, se recurren a variables topográficas, hidrológicas, geológicas, pedológicas y ambientales que influyen en diferente medida en la conductividad hidráulica subterránea y en la tasa de recarga de los acuíferos. Los resultados obtenidos indican que el mejor desempeño equiparable al estado del arte se obtiene para el modelo de bosques aleatorios (exactitud=0.77, puntaje F1=0.73, AUC=0.88) y que la construcción de modelos especializados en una región dada permite mejorar la capacidad de los modelos al reducir la varianza de los datos. Las variables más importantes en los modelos fueron: aspecto, densidad de drenaje, elevación, NDWI y precipitación. La principal limitación identificada en el desempeño de los modelos es la escasa cantidad y distribución irregular de los pozos de caudal conocido en el Perú, factor que parcializa el modelo hacia la costa, la región mejor documentada. El presente estudio sirve como marco referencial para la construcción de futuros modelos de aprendizaje automatizado una vez se amplíe el inventario público de pozos de agua subterránea o en caso privados introduzcan su propio inventario. El código empleado para el procesamiento de variables geoespaciales se encuentra en https://code.earthengine.google.com/fe63cd6184b009824ed3c843fdc5544d. El código utilizado para la construcción de modelos se encuentra registrado en Github en https://github.com/cesport/Tesis. Aplicaciones para visualizar los resultados de manera interactiva están disponibles para computadoras en https://cesarportocarrero.users.earthengine.app/view/gwp-peru y dispositivos móviles en https://cesarportocarrero.users.earthengine.app/view/gwp-peru-movil.

Agradecimientos

En primer lugar, quisiera agradecer a todos los que permitieron hacer posible el presente trabajo de investigación. Agradezco la mentoría de mi asesor Gustavo Larrea. Su apoyo constante, sus consejos, su conocimiento, su entusiasmo y su paciencia son factores que me permitieron seguir adelante y espero haber podido plasmar aunque sea una pizca en estas páginas. Inicialmente, yo elegí este tema de investigación entusiasmado únicamente por la posibilidad de incorporar metodologías novedosas y modernas en la ingeniería civil, sin mucho background teórico ni técnico. Sin su experiencia, todo el trabajo que se presenta en estas hojas no hubiera sido posible.

Asimismo, agradezco al profesor Ramzy Kahhat por su interés en el éxito de esta investigación. El profesor Kahhat estuvo presente desde el inicio del proceso, presentándome incluso a mi asesor de tesis. Sin embargo, diría que uno de los mayores aportes que recibí de parte suya fue mucho antes, cuando llevé el curso de Fundamentos de Ingeniería Ambiental, pues fue allí donde aprendí el valor de la investigación en nuestra disciplina.

Aprovechando que este trabajo marca el fin de toda mi formación universitaria, quisiera agradecer también a las personas que me permitieron llegar al final de este camino. A mis padres, Marco y Marina, que desde pequeño buscaron que siempre me esfuerce al máximo, sin olvidar soltar una carcajada aún en los peores momentos. A mi hermano, Marco, por contrastar tantas veces conmigo y brindarme siempre una perspectiva diferente en cada faceta de mi vida.

Finalmente, quisiera expresar un agradecimiento a todas las personas que hicieron de estos 5 años una experiencia inolvidable. A mis amigos de primer ciclo: Kail, Claudia y Carlos; que a pesar de los diferentes caminos que hemos tomado siempre estaremos el uno para el otro. A mis compañeros de intercambio en Alberta: CJ, Dong, Myles, Jiaming y Taka; que me hicieron sentir en casa aun estando en Canadá. A mis futuros colegas Lilian y Paolo, cuya amistad trascendió todos los estándares que uno podría imaginar.

A todos ustedes, muchas gracias.

Tabla de contenidos

Índice de figuras

Índice de tablas

Capítulo I

Introducción

1.1. Formulación y descripción del problema

Actualmente, la escasez de agua es un problema que aqueja al mundo y el Perú no es la excepción. En el contexto de consumo doméstico, se estima que menos del 70 % de la población del Perú dispone de acceso a la red de agua potable y que, en cinco departamentos, este porcentaje es inferior al 30 % (INEI, 2018). Tal es la magnitud del problema de la disponibilidad del agua en nuestro país que la ciudad de Lima se encuentra dentro de las 20 primeras ciudades con mayor estrés hídrico en el mundo (McDonald et al., 2014). Por otro lado, las consecuencias del estrés hídrico también han afectado al sector agrícola a tal punto de que el Ministro de Agricultura y Riego (Minagri) se vio obligado a invertir tres mil millones de soles en la perforación de pozos en los departamentos de Arequipa, Tacna, Moquegua y Puno (ANA, 2016). Sin embargo, en el norte, la dependencia en agua superficial ha conllevado a que exista escasez de agua a pesar de la presencia de los reservorios de Poechos y Gallito Ciego (Belletich, 2015).

En este sentido, una solución a los problemas de aprovisionamiento de agua consiste en promover la explotación de las reservas subterráneas de agua a nivel nacional, tal y como se está realizando en la costa sur. Esta propuesta resulta atractiva dado que, tan solo en la costa, existe una reserva explotable de casi 1600 millones de metros cúbicos. En otras palabras, estos pueden ser aprovechados sin comprometer la integridad de los acuíferos y del ecosistema (ANA, 2012). No obstante, resulta imprescindible primero profundizar el conocimiento que se tiene acerca del inventario hídrico subterráneo. Actualmente, no se dispone de datos acerca de la disponibilidad de agua subterránea de las vertientes hidrográficas del Amazonas y del Titicaca (ANA, 2012), lo cual evidencia el estado precario de la investigación hidrogeológica en el Perú.

Una de las principales consecuencias del escaso conocimiento de los recursos hídricos subterráneos nacionales tiene incidencia en el ámbito de la sostenibilidad. A pesar del alto potencial de agua subterránea presente en la costa como región, la explotación excesiva y desinformada de acuíferos ha provocado que los departamentos de Lambayeque, Tacna e Ica se encuentren en estado crítico (Guevara, 2014). Las consecuencias de este accionar han generado un descenso tal de la napa freática en la región de Ica que se ha atrofiado permanentemente el funcionamiento de los pozos. Se han reportado caudales menores de extracción, un deterioro de la calidad de agua recogida, y mayores costos de operación y mantenimiento en la explotación de acuíferos (Muñoz, 2015). La gravedad de este problema fue tal que los acuíferos del río Caplina, en Tacna; Villacuri en Ica y Motupe, en Lambayeque entraron en veda entre el 2006 al 2008 (Salazar, 2018). Asimismo, en el acuífero de La Yarada en Tacna, se ha registrado no solo una depresión considerable de la napa freática, sino también una reducción de la calidad del agua (Pino y Coarita, 2018).

Adicionalmente, cabe resaltar el impacto que tiene el desconocimiento de los recursos hídricos subterráneos en el aspecto económico. En ausencia de alternativas más económicas, para determinar si un pozo tiene un rendimiento adecuado para satisfacer las necesidades de los usuarios, es necesario efectuar perforaciones y estudios hidrogeológicos. Estos procedimientos suponen un costo considerable en términos de tiempo y dinero debido al uso de maquinaria pesada y mano de obra especializada. En este sentido, tanto empresas como gobiernos se ven forzados a invertir cuantiosas sumas de dinero para determinar cuáles son los emplazamientos en los que se puede extraer un mayor volumen de agua de manera sostenible. De no hacerlo, se construirían pozos con un caudal inferior al demandado o peor aún, sin presencia alguna de agua subterránea. Tal es el caso de la Municipalidad Distrital de Vista Alegre, en el cual el gobierno local invirtió un total de 5 millones de soles en un pozo donde el caudal de extracción era nulo (Huayta, 2018).

Una solución a la falta de información sobre los recursos hídricos subterráneos nacionales consiste en desarrollar un modelo de aprendizaje automatizado que clasifique al territorio peruano de acuerdo con su potencial de agua subterránea sin requerir de perforaciones. De este modo, durante la fase de anteproyecto, empresas y gobiernos podrían tener un referente adicional sin costo alguno para ubicar los emplazamientos en los cuales la construcción de pozos de alto caudal sería posible. Es necesario mencionar que este modelo no reemplaza a un estudio hidrogeológico realizado en campo. Sin embargo, sí puede ser empleado para efectuar una delimitación preliminar. Desde un punto de vista social, este producto permitiría conocer si un proyecto de infraestructura hidráulica es viable para aprovisionar de agua a comunidades aisladas y aliviar el estrés hídrico. Asimismo, desde una perspectiva de economía sostenible, esta iniciativa ayudaría principalmente a las industrias de uso intensivo de agua subterránea a construir pozos de alto rendimiento sin afectar la futura disponibilidad de este recurso.

En base a ello, la presente tesis de investigación desarrolló un modelo de clasificación de potencial de agua subterránea empleando dos metodologías de aprendizaje automatizado: bosques aleatorios y redes neuronales. Para el entrenamiento de estos modelos, se recurrió a información de la Autoridad Nacional del Agua (ANA) y a variables geoespaciales de índole topográfica, hidrológica, pedológica, geológica y ambiental. Cabe resaltar que la reducción de la inversión económica necesaria para la evaluación del agua subterránea es uno de los condicionantes principales de la presente investigación. En este sentido, todos los datos y métodos empleados son de acceso público, evitando así incurrir en costo alguno para la construcción de los modelos.

1.2. Preguntas de investigación

Pregunta de investigación general: ¿De qué manera se puede clasificar al territorio peruano de acuerdo con su potencial de agua subterránea relacionado a acuíferos de poca profundidad sin emplear perforaciones?

Preguntas de investigación específicas:

- ¿Cuáles son las variables topográficas, hidráulicas y geológicas que presentan una mayor correlación con el potencial de agua subterránea?
- ¿Cuál es el algoritmo de aprendizaje automatizado que permite efectuar una clasificación con mejor desempeño del potencial de agua subterránea?
- ¿Cómo debe segmentarse el territorio peruano para incrementar el desempeño de los modelos de clasificación de potencial de agua subterránea?

1.3. Objetivos

Objetivo general: Determinar el potencial hidrológico subterráneo asociado a acuíferos de poca profundidad en el Perú mediante la aplicación de modelos de clasificación empleando los algoritmos de aprendizaje de máquina de bosques aleatorios y redes neuronales artificiales.

Objetivos específicos:

- Determinar cuáles son las variables topográficas, hidrológicas, pedológicas, geológicas y/o ambientales que influyen con mayor preponderancia en el potencial de agua subterránea

- Comparar el desempeño de los modelos basados en los algoritmos de bosques aleatorios y redes neuronales para la clasificación del potencial de agua subterránea
- Evaluar si la segmentación por cercanía geográfica permite construir modelos de clasificación de agua subterránea con mejor desempeño

1.4. Hipótesis

Hipótesis general: Los modelos de aprendizaje automatizado sí son herramientas adecuadas para la aplicación de modelos de clasificación del potencial del agua subterránea en el Perú puesto que permiten abstraer fenómenos asociados a múltiples variables complejas en simultáneo, como es el caso de la hidrología subterránea.

Hipótesis específicas:

- Las variables más relevantes en la estimación del potencial de agua subterránea son: precipitación, evapotranspiración y tipo de suelo. Esto se debe a su influencia directa en la tasa de recarga de los acuíferos.
- El modelo implementado con redes neuronales artificiales presenta un mejor desempeño en comparación al de bosques aleatorios puesto que este presenta una mayor complejidad computacional.
- La segmentación por cercanía geográfica sí puede ser empleada para mejorar el desempeño de los modelos puesto que permite reducir la dispersión de las variables independientes y dependientes de acuerdo con la primera ley de la geografía de Tobler.

1.5. Justificación

Se estableció al agua subterránea como el objeto de estudio de esta investigación ya que es recurso de alto potencial en el Perú cuya documentación aún es sumamente precaria. Estudios del Programa de las Naciones Unidas para el Medio Ambiente (PNUMA) afirman que existe una deficiencia de estudios hidrogeológicos en el país para asegurar una explotación responsable de este recurso (Mendoza, Santayana, y Grover, 2010). En efecto, conocer más acerca del inventario hidrológico subterráneo resulta crucial para el desarrollo sostenible del país a nivel social y económico como se explica a continuación.

Desde un punto de vista social, el agua subterránea puede ser considerada como una herramienta clave para garantizar acceso al agua para toda la población. En el “Informe

Nacional sobre la Gestión del Agua en el Perú", la Comisión Económica para América Latina y el Caribe afirmó que: "El aprovechamiento de las aguas subterráneas ha recibido poca atención, debe ser objeto de un nuevo impulso para resolver el problema de las cuencas deficitarias" (Emanuel y Escurra, 2000, p. 9). Ciertamente, la rápida urbanización de las ciudades del Perú ha provocado que las fuentes de agua superficial sean insuficientes para abastecer a la población, especialmente en épocas de estiaje (Sedapal, 2017).

Desde un punto de vista económico, el incrementar la información acerca del agua subterránea para promover su uso resulta imprescindible para aterrizar los proyectos de accesibilidad de agua. Si bien en el pasado el gobierno ha invertido principalmente en la construcción de represas para garantizar un almacenamiento de agua que satisfaga a la demanda en crecimiento, el costo de estos proyectos es en promedio de $1.78 por metro cúbico. En cambio, el costo de un proyecto de pozo de agua subterránea es de $0.48 por metro cúbico (Rohde, 2014). Asimismo, el agua subterránea es menos vulnerable a contaminación superficial, pérdidas por evaporación y variaciones del ciclo de agua (Conagua, 2015). Por estos motivos, la inversión en pozos resulta potencialmente menos riesgosa y más económica que el agua superficial en algunas circunstancias.

De seguir manteniendo al agua subterránea en un estado de desconocimiento, la explotación sostenible de este recurso seguirá imposibilitada. Al no poder ubicar puntos de agua subterránea de alto rendimiento cerca de localidades alejadas sin acceso a agua, el Estado se verá forzado a recurrir únicamente a fuentes de agua superficial. La construcción de infraestructura que conecte a ríos y lagunas hasta los asentamientos urbanos puede llegar a costar hasta 37 veces más que un pozo de agua subterránea equivalente (Hierro, 2016). Por ejemplo, en comparación a la captación superficial, el bajo costo y plazo corto involucrados en el proceso constructivo de un pozo de extracción tubular han sido motivos por los cuales se ha considerado al agua subterránea como solución potencial al estrés hídrico en el departamento de Cajamarca (Santillán, 2018).

Además de complicar y encarecer la realización de obras de infraestructura hidráulica, la desinformación respecto a los recursos hídricos subterráneos también provocaría que prosiga el uso indebido e insostenible de este recurso. Ejemplo claro de ello se observa en la industria minera, donde la ubicación de lixiviados en zonas de alta infiltración y, por ende, alto rendimiento de agua subterránea, han afectado considerablemente la calidad de este recurso. Graves alteraciones del ecosistema han sido identificadas a causa de la acidificación y contaminación de los acuíferos aledaños a los centros mineros de Hualgayoc, Quiruvilca y Morococha (Tovar, 2005). Un plan de preservación de calidad de agua subterránea debería incluir una delimitación de zonas de amortiguamiento alrededor de los puntos con acuíferos de mayor potencial. En el caso particular de los acuíferos de poca profundidad, este punto

adquiere una mayor urgencia puesto que estos son los más susceptibles a ser contaminados.

Se eligió al aprendizaje de máquina como la metodología de trabajo adecuada para esta investigación debido a que es una rama de la inteligencia artificial con alto potencial para el desarrollo de modelos de regresión y clasificación. Si bien existen procedimientos alternativos al aprendizaje de máquina para el modelado de agua subterránea (ver Capítulo III) y no se puede garantizar la superioridad absoluta de una metodología sobre otra, los algoritmos de inteligencia artificial son capaces de manejar gran cantidad de datos en múltiples dimensiones y encontrar relaciones complejas y dinámicas entre las variables predictivas y la variable dependiente (Yang y Trewn, 2004). Asimismo, el uso del aprendizaje de máquina en el ámbito del agua subterránea ya ha sido explorado en países con deficiencia de fuentes de agua superficial tales como Zambia (Spiegel, 2017), Irán (Moghaddam et al., 2020; Naghibi, Pourghasemi, y Dixon, 2016; Rahmati, Pourghasemi, y Melesse, 2016) y Corea del Sur (Lee, Hong, y Jung, 2018; Lee, Song, Kim, y Park, 2012). A través del desarrollo de este modelo de clasificación para el territorio peruano, se busca contribuir al conocimiento hidrogeológico nacional, de modo que pueda ser empleado a futuro para proyectos que beneficien a la población y a la industria peruana. Cabe resaltar que la presente investigación se encuentra alineada con el Objetivo de Desarrollo Sostenible de las Naciones Unidas denominado “agua limpia y saneamiento”.

Capítulo II

Marco Teórico

2.1. Hidrología

Se entiende como la disciplina que estudia al comportamiento del agua en la Tierra en todas sus fases: sólida, líquida y gaseosa (Chow, 1988). Analiza el proceso de circulación del agua, así como sus propiedades fisicoquímicas y su interacción con los seres vivos. Dependiendo del enfoque espacial del estudio, esta disciplina se subdivide en hidrología superficial y subterránea.

2.1.1. Hidrología subterránea

Entiéndase como la rama de la hidrología que se encarga del estudio de la ocurrencia, distribución y movimiento de todos los cuerpos de agua debajo de la Tierra (Todd y Mays, 2005). En la especialidad de ingeniería civil, se emplea frecuentemente la hidrología subterránea para el diseño de pozos de explotación de acuíferos, obras de drenaje y el control de la napa freática en proyectos de construcción.

2.1.2. Pozos de agua subterránea

Dícese de todo orificio vertical excavado en la tierra para la extracción de agua proveniente de un acuífero libre o confinado. Adicionalmente, se pueden emplear para propósitos de exploración subterránea, recarga artificial de acuíferos y deposición de aguas servidas (Todd y Mays, 2005). Una de las características más importantes de estos elementos

corresponde a su rendimiento o caudal seguro, este se define como la tasa máxima permisible de extracción. De exceder este valor, se puede dañar irreversiblemente al acuífero (Campillo, 2010). Teóricamente, el máximo caudal de extracción en un pozo en un acuífero libre se puede modelar empelando la expresión planteada por Driscoll (1987) con los resultados de dos piezómetros. Esta se presenta en la ecuación 2.1.

$$Q = \frac{1.366K(H^2 - h^2)}{\log(R/r)} \tag{2.1}$$

Donde K es la conductividad hidráulica del acuífero, H y h corresponden a los tirantes medidos con los piezómetros y R y r son las distancias de estos instrumentos hacia el centro del pozo. Pese a la popularidad de esta ecuación, se cuestiona altamente la idoneidad de esta puesto que se basa en la caracterización de Thiem (1906) para el flujo subterráneo. En otras palabras, se asume que el flujo de agua es únicamente horizontal-radial y que se da en un acuífero sin desniveles e isotrópico. Asimismo, la fórmula asume que el pozo analizado penetra la totalidad del espesor del acuífero y que la recarga solo se efectúa en la periferia del cono de depresión. Si bien estas limitaciones eran consideradas aceptables en la época que esta ecuación fue formulada hace 100 años, se ha demostrado que las presuposiciones impiden modelar adecuadamente el comportamiento en los pozos al ignorar múltiples fuentes de recarga y pérdidas (Tügel, Houben, y Graf, 2016). Por este motivo, actualmente se prefiere la ejecución de ensayos in situ con bombas para la determinación del caudal de explotación de los pozos de agua subterránea.

2.1.3. Potencial de agua subterránea

Se define como una propiedad geoespacial asociada a un emplazamiento. Este indica cuánto sería el rendimiento hipotético de un pozo de agua subterránea de ser construido en dicha ubicación. Este valor depende directamente de la tasa de recarga del acuífero, los mecanismos de recarga, su capacidad de almacenamiento y las propiedades de transmisibilidad en el subsuelo (Kebede, 2013). Todas las propiedades antes mencionadas dependen a su vez de múltiples variables topográficas, hidrológicas y geológicas cuya relación es sumamente compleja.

2.2. Aprendizaje de máquina

Se entiende como una rama de la inteligencia artificial dedicada a la detección automatizada de patrones relevantes en grandes paquetes de información. Ejemplos de aplicaciones en este rubro incluyen la implementación de filtros automáticos de correos electrónicos, reconocimiento vocal, procesamiento de imágenes y una gran variedad de predicciones (meteorológicas, financieras, etc.). En el ámbito de la hidrogeología, intervienen en simultáneo múltiples variables: litología, relieve, topografía, porosidad, estructuras geológicas, patrones de drenaje, evapotranspiración, precipitación, entre otros (Rathay, Allen, y Kirste, 2018). Debido a la complejidad del problema, modelos numéricos tradicionales no han demostrado ser adecuados para modelar el comportamiento del flujo del agua subterránea en los pozos (Shenga, Baroková, y Šoltész, 2018). En cambio, el aprendizaje de máquina permite generar modelos capaces de manejar variables con relaciones complejas (no lineales, no monotónicas y multimodales) sin efectuar presuposiciones sobre los datos como es típico en modelos paramétricos tradicionales (Olden, Lawler, y Poff, 2008).

En líneas generales, en la elaboración de modelos de aprendizaje automatizado se distinguen tres fases: entrenamiento, validación y evaluación. Los *datos de aprendizaje*, es decir, datos individuales de variables independientes asociadas con su respectiva variable dependiente, se dividen también en tres grupos para ser utilizados en estas tres fases (Louppe, 2014). En primer lugar, el modelo se construye en base a los datos de entrenamiento y se calculan las métricas de desempeño para los datos de validación. Luego, las características del modelo (denominadas *hiperparámetros*) se actualizan iterativamente para maximizar el desempeño del modelo con los datos de validación. Finalmente, con los *hiperparámetros* definitivos, se calcula el desempeño del modelo utilizando los datos de evaluación. Puesto que el modelo no ha sido expuesto a este conjunto de datos, las métricas obtenidas en esta etapa corresponden al puntaje final del modelo construido.

2.2.1. Validación cruzada

Se define como una técnica empleada durante las etapas de entrenamiento y validación para evitar que el modelo no pierda su capacidad de generalización. En vez de tener dos grupos estáticos de datos de entrenamiento y validación, los datos se dividen en n grupos donde n es un valor seleccionado por el usuario. Luego, en una primera iteración, el primer grupo creado se emplea para validación del modelo mientras que los $n-1$ restantes se emplean para el entrenamiento. En una segunda iteración, el segundo grupo asume el rol de datos de validación y nuevamente los $n-1$ grupos restantes se usan para el entrenamiento.

Este proceso se efectúa sucesivamente *n* veces. Finalmente, para evaluar el desempeño de validación, se emplea el promedio de las métricas obtenidas en las *n* iteraciones. Una representación gráfica del procedimiento de validación cruzada se presenta en la figura 2.1.

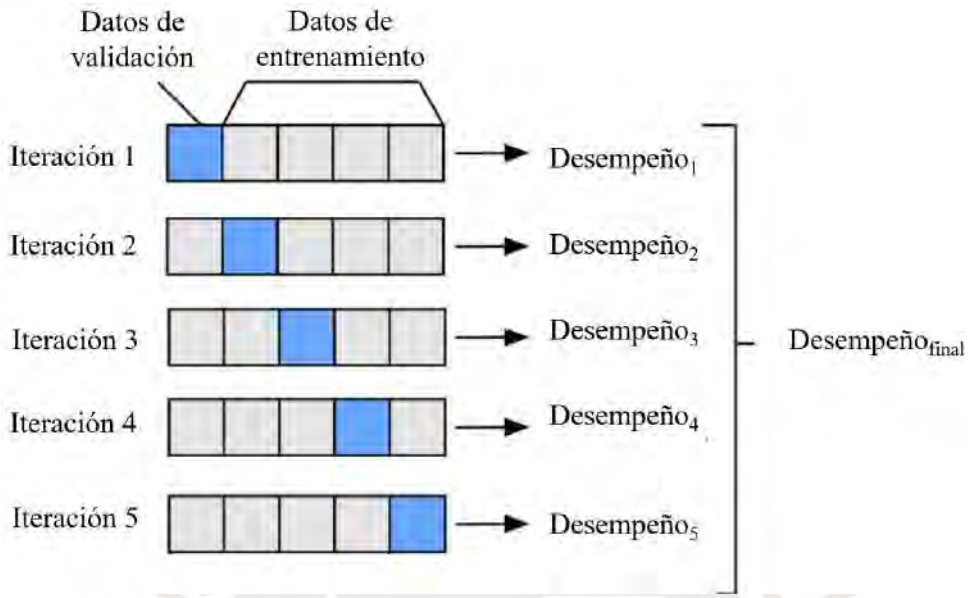


Figura 2.1. Ejemplo de validación cruzada empleando 5 iteraciones.

2.2.2. Bosques aleatorios

Entiéndase como un algoritmo empleado como método de clasificación y regresión a través del uso de múltiples árboles de decisión. Estos, a su vez, definen como diagramas en los cuales cada división se realiza en base a decisiones jerarquizadas. Cada decisión se efectúa en un "nodo" y está asociada a cada una de las variables independientes empleadas en el modelo. Haciendo analogía a un árbol, se le conoce al nodo inicial como "raíz", y a los inferiores como "hojas". En la figura 2.2, se ejemplifica un árbol de decisión simplificado para el presente caso de estudio, por lo cual la variable dependiente es el potencial de agua subterránea representado en tres posibles categorías: bajo, moderado y alto. En este ejemplo, se abordan como variables independientes únicamente a la distancia a ríos y al tipo de relieve. Se puede apreciar que, para cada emplazamiento, luego de recorrer los nodos correspondientes de acuerdo con sus propiedades, se puede asignar una clasificación del potencial de agua subterránea. Para determinar cómo se efectúan las divisiones en cada nivel del árbol de decisión, se busca emplear variables que permitan segmentar a los datos en grupos cada vez más puros, es decir, correspondientes a una misma clase. Para medir el grado de pureza, se pueden emplear dos criterios principalmente: el índice de Gini o la entropía de la información.

Figura 2.2. Ejemplo simplificado de árbol de decisión para el caso de estudio.

Índice de Gini

Criterio asociado al algoritmo CART (Breiman, Friedman, Stone, y Olshen, 1984), este parámetro se define en la ecuación 2.2.

$$S = 1 - \sum_{i=1}^{n} P_i^2 \tag{2.2}$$

Donde P_i es la probabilidad de que un elemento pertenezca a la categoría iésima. De manera intuitiva, se interpreta al índice de Gini como la probabilidad de que un elemento sea clasificado incorrectamente si fueran clasificados aleatoriamente de acuerdo con la distribución de clases existente. Un valor de 0 para el índice de Gini supone sería de 0 representa pureza total, es decir, un grupo de datos de una única clase. Por otra parte, si los datos se dividieran perfectamente en dos clases de igual tamaño, se obtendría un valor de 0.5 para este parámetro (impureza perfecta para el caso de dos categorías).

Entropía de la información

Criterio asociado a los algoritmos ID3 y C4.5. Se desarrolló ID3 con el objetivo de construir árboles de decisión en base a variables categóricas (Quinlan, 1986). C4.5 es una extensión de ID3, pero su utilidad se extiende a variables numéricas. La formulación matemática de la entropía se presenta en la ecuación 2.3.

$$S = -\sum_{i} P_i log_2(P_i) \tag{2.3}$$

Donde P_i nuevamente es la probabilidad de que un elemento pertenezca a la categoría iésima. A diferencia del índice de Gini, un conjunto perfectamente impuro para un problema de clasificación de dos clases presenta un valor de 1.

Uso de múltiples árboles

En un bosque aleatorio, cada árbol de decisión comprende un subconjunto de datos y de variables independientes, generando varios modelos no interdependientes. La categoría final asignada por un modelo de este tipo se obtiene calculando el promedio de los resultados obtenidos en cada árbol. De este modo, se logra evitar que los modelos sobreentrenen y se parcialicen hacia los datos de entrenamiento. En otras palabras, el desempeño del modelo aún es adecuado cuando se enfrente a problemas de clasificación de datos inéditos. Un bosquejo típico de bosque aleatorio simplificado para la presente investigación se puede observar en la figura 2.3. Se puede apreciar que como la mayoría de los árboles indican que un determinado emplazamiento presenta un alto potencial de agua subterránea, razón por la cual esta es la clasificación final asignada.

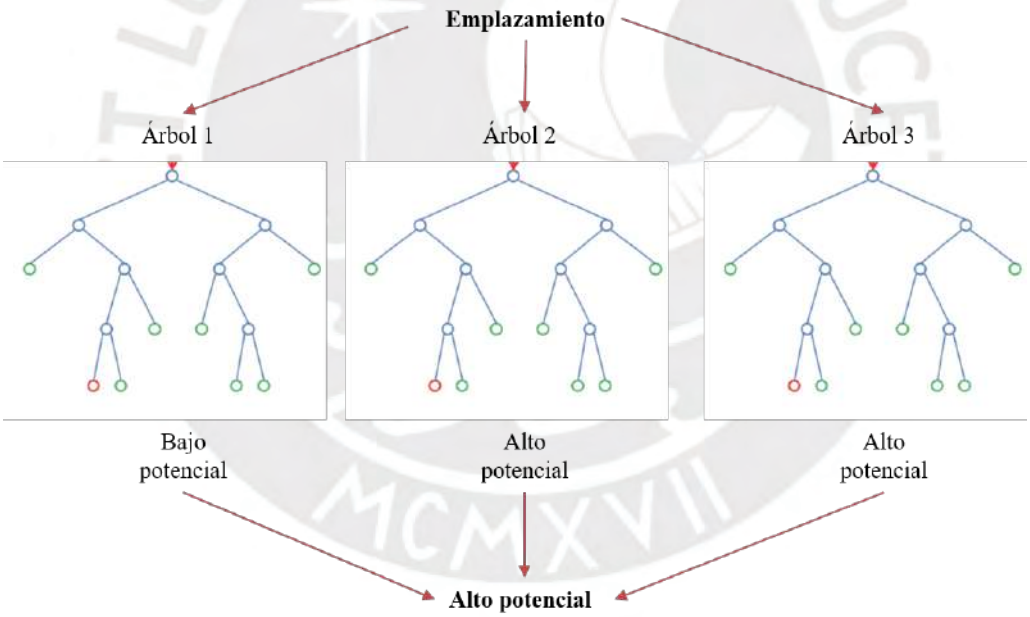


Figura 2.3. Ejemplo simplificado de bosque aleatorio con tres árboles de decisión para el caso de estudio.

2.2.3. Redes neuronales artificiales

Entiéndase como un sistema basado en componentes electrónicos o software que busca desarrollar una función de interés simulando el comportamiento del cerebro humano. Una de las aplicaciones más conocidas consiste en la simulación de la capacidad de aprendizaje humano para la aplicación de modelos de clasificación y regresión (Haykin et al., 2009).

Elementos

La mínima unidad funcional en las redes neuronales artificiales son las neuronas. Como se observa en la figura 2.4, cada neurona se encarga de recibir un número determinado de valores de entrada (x_i) de otras neuronas y los multiplican por pesos distintos (ω_i). Luego, estos valores son sumados junto a un factor adicional denominado sesgo (b). Finalmente, a este sumando se le aplica una función de activación (φ) para obtener el valor de salida (y). Dependiendo de las necesidades del proyecto, se puede elegir como función de activación a la función sigmoide, tangente hiperbólica, lineal rectificada, entre otras. Cabe resaltar que el comportamiento de las neuronas artificiales se diseñó para imitar a las neuronas humanas (Guresen y Kayakutlu, 2011). Una neurona biológica recibe estímulos (valores de entrada) y los conecta hacia otras neuronas a través de la sinapsis (pesos) entre ellas. Luego, los estímulos son acumulados en las dendritas (función sumatoria) y modificados por el potencial de acción (sesgo) para regular la propagación de señales eléctricas en el sistema nervioso. Finalmente, el soma o cuerpo celular de la neurona (función de activación) transforma el estímulo recibido antes de transmitirlo al axón (valor de salida) para distribuirlo a otras neuronas.

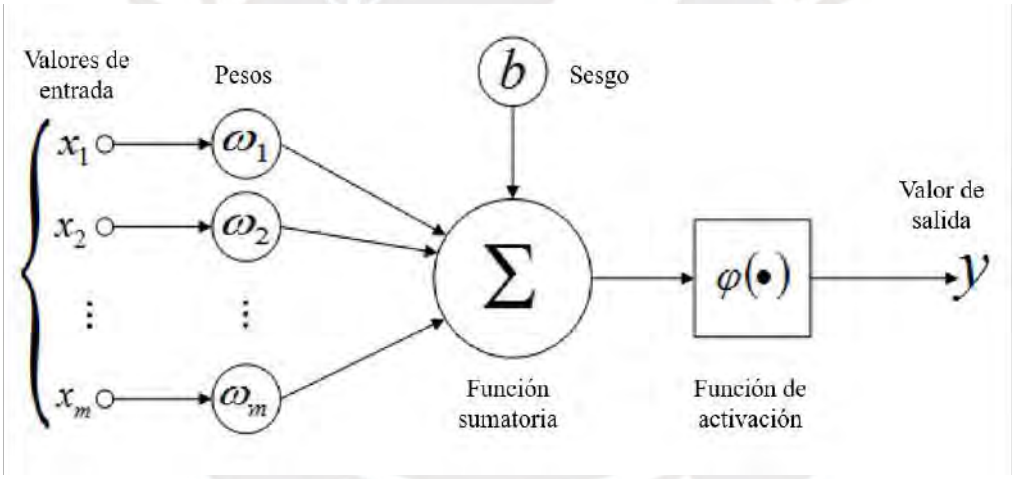


Figura 2.4. Elementos de una neurona típica en una red neuronal artificial.

En un modelo de redes neuronales artificiales las neuronas se organizan en capas. Estas se clasifican dependiendo de su ubicación como se presenta en la figura 2.5. La descripción de cada categoría se presenta a continuación:

- Capa de entrada

  Entiéndase como la capa de neuronas ubicada al inicio de la red neuronal. Esta es comparable a los receptores nerviosos en una red neuronal biológica, órganos cuya labor consiste en recibir los estímulos provenientes del ambiente. En el caso de las redes neuronales artificiales, las neuronas de la capa de entrada se encargan de recibir los valores de las variables independientes.

- Capa de salida

  Dícese del conjunto de neuronas ubicado al final del modelo de la red neuronal artificial. En un modelo de regresión, estas se encargan de computar el valor predicho. En el caso de un modelo de clasificación, su labor consiste en determinar la probabilidad de correspondencia de un elemento a una o varias categorías. En comparación a una red biológica, esta capa presenta un comportamiento similar a los efectores nerviosos, es decir, órganos como músculos y glándulas que producen una respuesta frente a un estímulo.

- Capas ocultas

  Elementos ubicados entre las capas de entrada y salida. Se puede emplear una o múltiples capas ocultas para modelar relaciones complejas no lineales entre las variables independientes y dependiente. Su símil en el ámbito biológico corresponde a la red de neuronas que se encuentra entre los receptores y efectores.

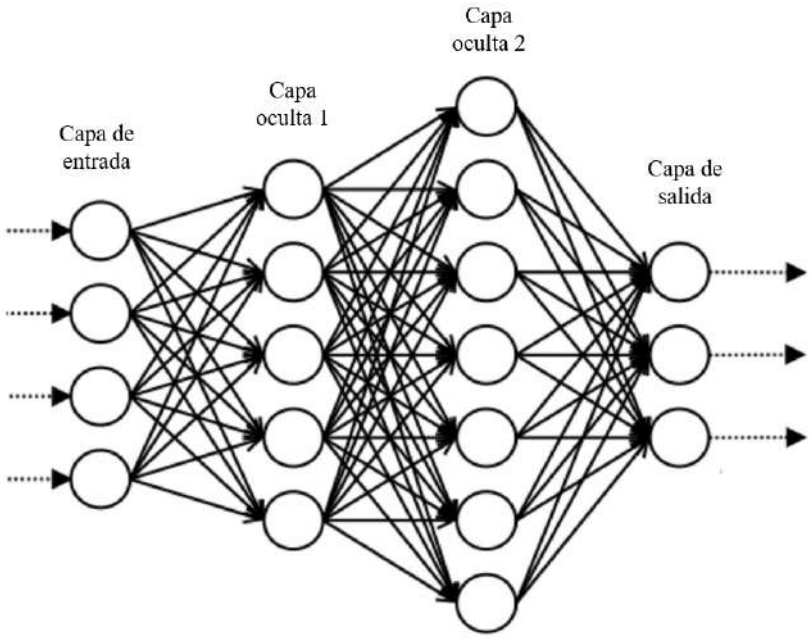


Figura 2.5. Ejemplo de arquitectura de una red neuronal artificial con dos capas ocultas.

Función costo y optimizadores

Los valores asociados a los pesos y a los sesgos en una red neuronal artificial se inicializan aleatoriamente. Sin embargo, estos valores se modifican constantemente a lo largo del entrenamiento para minimizar el error de predicción o clasificación del modelo. Este, se cuantifica a través de una función denominada función de pérdida, cuya elección depende del tipo de problema. En el caso de modelos de regresión, posibles funciones de pérdida incluyen al error promedio cuadrado, error medio absoluto, etc. Por otro lado, para problemas de

clasificación se suele recurrir a la función de entropía cruzada categórica. Para reducir el valor de la función de pérdida y, por ende, el error del modelo, se emplean optimizadores para actualizar los pesos y sesgos en la red neuronal. Si bien existen múltiples optimizadores, la gran mayoría involucra al descenso de gradiente estocástico. Este proceso busca utiliza las gradientes de la función costo para hallar la solución que converja en el mínimo global de la función costo. Puesto que el cálculo de las gradientes de la función costo se inicia en las neuronas de la capa de salida y termina en la capa de entrada empleando derivación con la regla de la cadena, se afirma que este proceso se realiza por retro-propagación.

Parámetros de entrenamiento

Cuando un modelo se entrena en base a un conjunto de datos, se debe especificar el número de épocas (*epochs*) y el tamaño del lote (*batch size*). El número de épocas se entiende como el número de veces que los datos son utilizados en su totalidad. Por otro lado, el tamaño de lote se define como el número de datos empleados para generar particiones de la base de datos original. Al culminar de procesar un lote, el modelo actualiza internamente sus pesos y sesgos para mejorar su desempeño. Por ejemplo, si se tiene una base de datos de 100 elementos y se define un número de épocas de 2 y un tamaño de lote de 50, se estaría programando el modelo de redes neuronales artificiales para que optimice sus parámetros internos cada 50 elementos. Asimismo, puesto que se está revisando los 100 elementos de la base de datos dos veces, se estarían actualizando los valores de pesos y sesgos de neuronas un total de 4 veces.

Regularizadores

Los modelos de redes neuronales aprenden con facilidad a reconocer patrones por lo que son muy susceptibles a sobreentrenarse, perdiendo generalización. En otras palabras, tienden a reducir su desempeño al exponerlos a nuevos datos (Liška, Kruszewski, y Baroni, 2018). Para subsanar este impase, se recurren a regularizadores de tipo L1, L2 y de descarte (*dropout*). De manera general, el proceso de regularización consiste en penalizar algunas neuronas con el objetivo de reducir la complejidad del modelo y evitar la parcialización de este. En el caso de L1 y L2, el proceso de regularización consiste en sumar un factor adicional la función de pérdida para penalizar a una neurona. La diferencia entre ambos se basa en la magnitud de este factor. Mientras que en el caso de L1 el factor es proporcional al valor absoluto del peso de las neuronas, en L2 es proporcional al cuadrado del peso. En términos prácticos, esto provoca que en el caso de L1 el peso de las neuronas tienda a cero durante el proceso de optimización. Gracias a ello, este resulta más útil en problemas donde intervengan una gran cantidad de variables independientes donde se requiera purgar las menos importantes para reducir la complejidad del modelo. Por otro lado, la regularización por medio de descarte consiste en asignar aleatoriamente pesos de 0 en un porcentaje definido de neuronas.

Capítulo III

Estado del Arte

3.1. Procedimientos in situ para la medición del potencial de agua subterránea

3.1.1. Medición directa: ensayo de bombeo de caudal escalonado

Históricamente, el estudio del potencial de agua subterránea se ha efectuado mediante procedimientos geofísicos e hidrogeológicos in situ. Actualmente, el método más conocido para determinar el caudal máximo explotable de un pozo es el ensayo de bombeo de caudal escalonado o caudal variable creado por Jacob (1947). Para realizar este procedimiento, el primer paso consiste en la perforación del suelo y la construcción del pozo. Esta técnica requiere de excavadoras, barrenas, anillos de concreto y filtros de grava (Gestmontes, 2010). El procedimiento consiste en extraer el agua del pozo durante una hora a caudal constante con un equipo de bombeo y medir el descenso correspondiente del nivel del agua (abatimiento) al final del intervalo. Posteriormente, se incrementa el caudal y se repite la experiencia durante otro periodo de una hora, se registra el abatimiento y se repite el procedimiento sucesivamente (Torres, 2006). La elección de los valores de caudal empleados para el ensayo debe caracterizar adecuadamente el régimen del pozo, por lo cual debe ser realizada previamente por un experto.

Determinación del caudal seguro

Según Jacob (1947), el abatimiento en un pozo se puede aproximar matemáticamente por la suma de dos monomios, el primero proporcional al caudal de extracción y el segundo

proporcional al cuadrado de este. La evidencia empírica indica que, para caudales bajos, prima el comportamiento lineal de abatimiento. No obstante, para caudales altos el monomio de comportamiento cuadrático adquiere importancia y eleva considerablemente los niveles de abatimiento. Caudales de extracción en este rango provocan el colapso del pozo y pérdidas adicionales debido a la formación de un flujo turbulento. El caudal límite a partir del cual se presenta este comportamiento se denomina caudal seguro o rendimiento máximo del pozo y se determina empleando un método gráfico. Primero, se grafican los caudales de bombeo en el eje de abscisas con su abatimiento respectivo en el eje de ordenadas. Seguidamente, se trazan dos rectas de ajuste: una para el tramo lineal inicial y una para el tramo no lineal final. La intersección de ambas corresponde al punto de inflexión aproximado para el comportamiento del abatimiento. Por este motivo, el caudal asociado a este punto de inflexión corresponde al rendimiento del pozo (Van Tonder, Botha, y Van Bosch, 2001). De exceder este valor por un tiempo prolongado, se presentan daños irreparables el pozo y se espera una reducción gradual de su caudal hasta su agotamiento (Campillo, 2010). A modo de ejemplo, se graficaron los resultados obtenidos por Jimenez (2017) en el ensayo de caudal escalonado realizado a un pozo empleado por el fundo agroindustrial “La Punta” ubicado en la localidad de Huaura. Como se puede apreciar, las rectas de ajuste correspondientes a los comportamientos lineal y no lineal de abatimiento coinciden en un punto de intersección cuyo caudal asociado es de aproximadamente 49 lps, valor que fue redondeado a 50 lps por la investigadora.

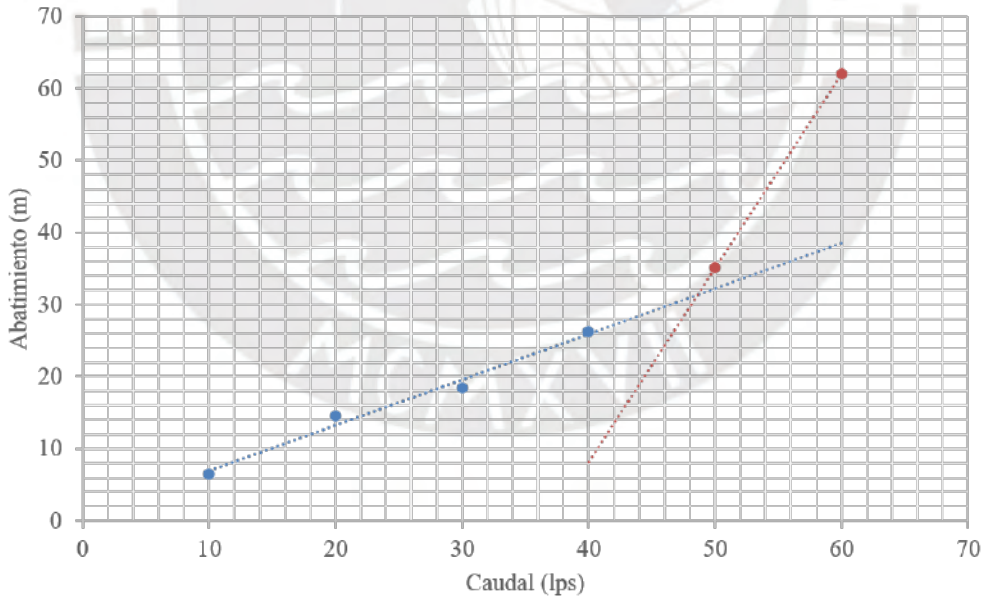


Figura 3.1. Ejemplo de determinación del caudal seguro de un pozo mediante el ensayo de bombeo de caudal escalonado. Graficado en base a los resultados de Jimenez (2017).

3.1.2. Medición indirecta: ensayo de resistividad eléctrica

En el ámbito global, una primera propuesta para estimar el rendimiento de pozos en la fase de anteproyecto se relaciona con la medición de la resistividad eléctrica del suelo. Este procedimiento involucra el uso del método de cuatro electrodos de Wenner para determinar la resistencia en ohmios asociada a una superficie. En el caso de áreas de estratos rocosos, los puntos con mayor potencial de agua subterránea se encuentran asociados a las zonas con menor resistencia al flujo eléctrico (AGI, 2018). Esto se debe a que la baja resistividad está asociada a la ocurrencia de fracturas en la roca. En estas zonas fluyen grandes cantidades de agua con minerales disueltos, solución que se caracteriza por tener una menor resistividad que la roca. Por otro lado, en el caso de suelos compuestos principalmente por sedimentos, la determinación de zonas con alto rendimiento de agua subterránea se relaciona con una alta resistividad eléctrica. Esta diferencia se produce puesto que los suelos arcillosos son mejores conductores eléctricos que los suelos gravosos y arenosos. Dado que la arcilla está compuesta por partículas finas y dificulta el paso del agua, es en cierta medida impermeable e inadecuada para la construcción de pozos. Empleando levantamientos de resistividad eléctrica, se logró cuantificar el rendimiento de los pozos en las áreas cercanas al Himalaya (Israil, Al-hadithi, y Singhal, 2006) y en la región de Bengala (Jha, Chowdary, y Chowdhury, 2010).

3.2. Procedimientos ex situ para la medición del potencial de agua subterránea

La principal limitación identificada en los métodos antes mencionados se relaciona con la necesidad de efectuar el levantamiento de datos de caudal o resistividad en toda el área de estudio empleando equipos (bombas, barrenos, electrodos, entre otros) y mano de obra capacitada. En este sentido, para caracterizar el potencial de agua subterránea de toda una región o un país, sería necesario efectuar múltiples levantamientos durante varios días, lo cual no sería factible en términos económicos. Por este motivo, la hidrogeología recurrió al uso de variables geoespaciales medidas de manera remota para la aplicación de modelos de clasificación del potencial de agua subterránea a nivel macro. Para la ejecución de todos estos métodos, se denomina variables independientes a las propiedades topográficas, hidrológicas y geológicas empleadas para estimar la variable dependiente, que en este caso hace referencia al potencial de agua subterránea. Estos procedimientos se emplean principalmente para la caracterización del potencial asociado a acuíferos libres y semiconfinados puesto que las variables empleadas describen propiedades superficiales o de baja profundidad.

3.2.1. Modelos dependientes de un panel de expertos

Estos modelos de clasificación del potencial de agua subterránea se desarrollaron empleando dos técnicas: el Análisis de Decisión Multicriterio (ADM) y el Proceso de Jerarquía Analítica (PJA). Ambos modelos utilizan factores de importancia para otorgar un peso a múltiples variables independientes que condicionan el rendimiento de un pozo. La suma de todos los ponderados permite determinar si posee alto o bajo rendimiento de explotación. Para la determinación de los pesos asociados a cada una de las variables independientes, se recurre a un comité de expertos que en base a su experiencia otorgan mayor o menor importancia a cada factor. La diferencia principal entre ambos modelos es que en la metodología PJA se puede establecer un límite de inconsistencia entre la opinión de los expertos mientras que el ADM no presenta este tipo de restricción. Modelos PJA para mapear el potencial de agua subterránea fueron empleados en India para la cuenca del río Vaigai (Kaliraj, Chandrasekar, y Magesh, 2014), en la región de Bengal (Chowdhury, Jha, Chowdary, y Mal, 2008) y en Udaipu (Machiwal, Jha, y Mal, 2011). Por otro lado, modelos ADM fueron utilizados en Oke-Ana, Nigeria (Akinlalu, Adegbuyiro, Adiat, Akeredolu, y Lateef, 2017) y en Kedah, Malasia (Adiat, Nawawi, y Abdullah, 2012). Sin embargo, cabe resaltar que la principal limitación de estos métodos es la dependencia en la opinión de terceros. Este factor condiciona los resultados de la investigación al juicio de un grupo de personas, sesgando el desempeño de los modelos. Una alternativa para evitar este sesgo consiste en emplear modelos paramétricos y no paramétricos, completamente independientes de la opinión de las personas.

3.2.2. Modelos independientes de un panel de expertos

Estadística tradicional: modelos paramétricos

Estas metodologías se construyen en base a modelos paramétricos estadísticos que permiten caracterizar el comportamiento de dos variables (bivariados) o múltiples variables en simultáneo (multivariado) en función de una distribución de probabilidad. Existen múltiples instancias de modelos paramétricos de clasificación del potencial de agua subterránea en la literatura: regresión logística para la región de Akesehir, Turquía (Ozdemir, 2011); regresión lineal generalizada y aditiva para Lorestán, Irán (Falah, Ghorbani Nejad, Rahmati, Daneshfar, y Zeinivand, 2017); tasa de frecuencia probabilística para Pohrang, Corea del Sur (Oh, Kim, Choi, Park, y Lee, 2011); función de creencia probatoria para Langat, Malasia (Nampak, Pradhan, y Manap, 2014); pesos de evidencia para Birjand, Irán (Pourtaghi y Pourghasemi, 2014) y factor de certeza para Varamín, Irán (Razandi, Pourghasemi, Neisani, y Rahmati, 2015). A pesar del nivel de éxito que obtuvieron estos modelos, la principal limitación de

estos se relaciona con todas las presuposiciones que se tienen que efectuar a priori para su funcionamiento. Estas condicionan forzosamente el comportamiento de las variables independientes y dependientes, las cuales en realidad poseen distribuciones erráticas al estar asociadas a la hidrogeología y el medio ambiente. Por ejemplo, en el caso de un modelo de regresión logística, se asume que las variables presentan una distribución normal y que existe homocedasticidad en el modelo, es decir que existe una varianza constante en el error de todas las observaciones (Kleinbaum y Klein, 2010). Ambos condicionantes no se pueden garantizar al tratar con variables de la índole de precipitación, elevación, tipo de suelo, entre otras que se relacionan con el potencial de agua subterránea. Asimismo, todos los modelos paramétricos asumen que las variables independientes no presentan relación mutua alguna, factor cuestionable al tratar con propiedades hidrológicas, geológicas y topográficas.

Aprendizaje automatizado: modelos no paramétricos

Para solucionar las limitaciones de los modelos predictivos tradicionales, se recurrieron a algoritmos de inteligencia artificial. El modelado algorítmico no asume ningún comportamiento para las variables independientes y dependiente, por ende, no se elige ninguna distribución de probabilidades ni es necesario efectuar pruebas de bondad de ajuste, análisis de residuos, etc. Este enfoque predictivo prioriza efectuar la estimación en base a la observación de patrones en los datos de entrada, sea cual sea la relación entre las variables independientes y la variable dependiente (lineal, exponencial, logarítmica, entre otras). Por este motivo, estos modelos no presentan contratiempos al enfrentarse a grandes cantidades de datos y dimensiones (Breiman, 2001). Gracias a esta ventaja, son una herramienta sumamente útil en la predicción de fenómenos donde intervienen grandes cantidades de variables independientes. Tal es el caso del agua subterránea, cuyo caudal se ve afectado simultáneamente por variables topográficas, hidrológicas y geológicas, entre otras.

Existen diversos algoritmos de aprendizaje automatizado que se han empleado para caracterizar el potencial de agua subterránea en los últimos años. Por ejemplo, en Irán, se han empleado máquinas de soporte vectorial para la provincia de Ardebil (Naghibi, Ahmadi, y Daneshi, 2017) y modelos de entropía máxima para Mehrán (Rahmati et al., 2016). Sin embargo, las métricas más altas de desempeño se han reportado en modelos basados en los algoritmos de bosques aleatorios y redes neuronales, razón por la cual estos fueron seleccionados para caracterizar el potencial de agua subterránea en la presente investigación. En la tabla 3.1 se presenta un resumen de las últimas investigaciones de potencial de agua subterránea en las cuales se emplearon estos algoritmos.

Tabla 3.1. Características de los modelos de agua subterránea basados en bosques aleatorios y redes neuronales correspondientes al estado del arte.

Autor	Área de estudio	Variables independientes	Modelo	Desempeño
Lee et al. (2012)	Pohang, Corea del Sur	Elevación Elevación media en un radio de 300m Elevación media dentro de la cuenca Pendiente Pendiente media entro de la cuenca Curvatura del terreno Índice topográfico de humedad Densidad de ríos Área de cuenca Densidad de longitud de fallas Densidad de frecuencia de fallas Unidades geológicas Textura del suelo	Redes neuronales artificiales	AUC[1]: 0.77
Naghibi et al. (2016)	Koohrang, Irán	Pendiente Elevación Aspecto topográfico índice topográfico de humedad Longitud de pendiente Curvatura de perfil Curvatura plana Distancia a ríos Distancia a fallas Densidad de drenaje Densidad de fallas Litología Uso de tierras	Bosques aleatorios	AUC[1]: 0.71
Rahmati et al. (2016)	Mehrán, Irán	Elevación Pendiente Aspecto topográfico Curvatura plana Distancia a ríos Densidad de drenaje Índice topográfico de humedad Uso de tierras Litología Textura del suelo	Bosques aleatorios	AUC[1]: 0.83
Spiegel (2017)	Zambia	Tasa de recarga de acuíferos Conductividad hidráulica Relieve Densidad de drenaje	Bosques aleatorios	Exactitud[2]: 0.58

Tabla 3.1 (cont.). Características de los modelos de agua subterránea basados en bosques aleatorios y redes neuronales

Autor	Área de estudio	Variables independientes	Modelo	Desempeño
Lee et al. (2018)	Boryeong, Corea del Sur	Elevación Elevación media en un radio de 300m Pendiente Pendiente media en un radio de 300m Índice de fuerza de corriente Área de la cuenca Densidad de ríos Distancia a ríos Formaciones geológicas Litología Densidad de longitud de fallas Densidad de frecuencia de fallas Puntos de intersección de fallas Tipo de suelo Densidad forestal Profundidad de agua subterránea Gradiente hidráulico	Redes neuronales artificiales	AUC[1]: 0.84
Moghaddam et al. (2020)	Hableh-Roud, Irán	Índice topográfico de humedad Posición relativa de la pendiente Curvatura plana Curvatura en perfil Índice topográfico de posición Índice de aspereza del terreno Distancia a fallas Densidad de drenaje	Bosques aleatorios Redes neuronales artificiales	AUC[1]: (BA) 0.94 AUC[1]: (RNA) 0.87

[1] Área debajo de la curva de Características Operativas de Receptor, métrica que evalúa tanto la sensibilidad y la especificidad del modelo independientemente del umbral de decisión.
[2] Cociente entre el número de elementos correctamente clasificados y el total de elementos.

Capítulo IV

Metodología y materiales

En líneas generales, el procedimiento investigativo se dividió en tres etapas: preparación de la base de datos, construcción y evaluación de modelos de clasificación, y generalización de modelos de clasificación. Un diagrama que resume la metodología abordada se presenta en la figura 4.1.

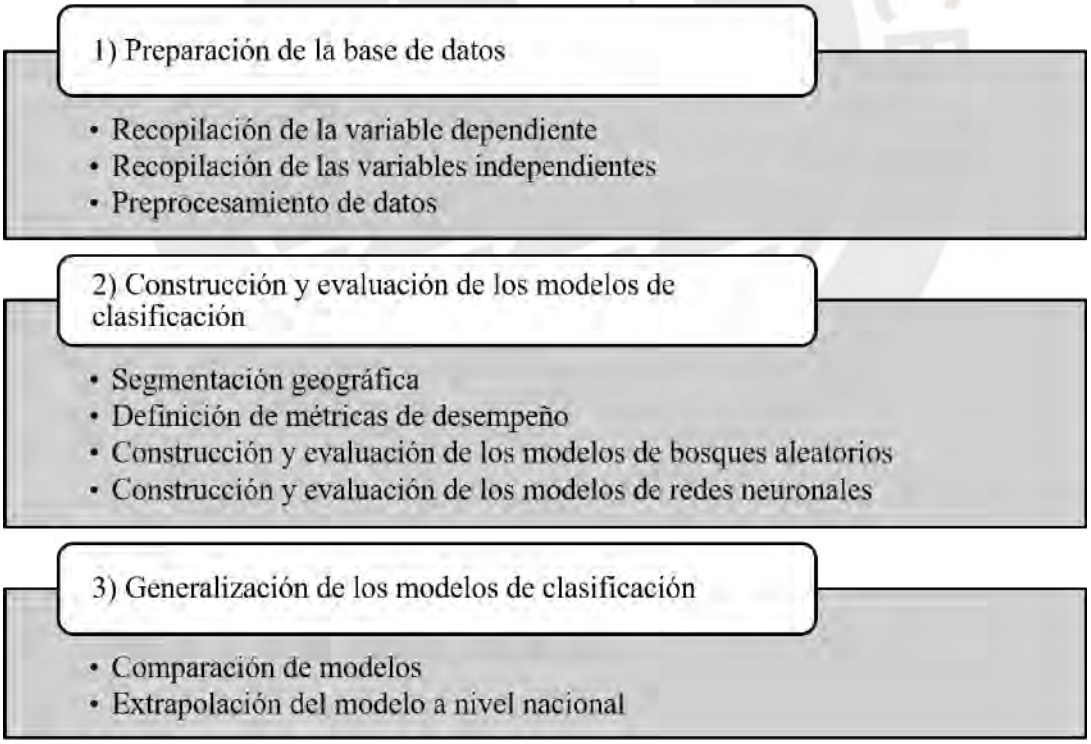


Figura 4.1. Diagrama del proceso metodológico abordado en la presente investigación.

4.1. Preparación de la base de datos

4.1.1. Recopilación de la variable dependiente

Selección de pozos con caudal conocido

Se recopilaron los informes del Minagri relacionados con el inventariado, monitoreo y evaluación de los pozos de agua subterránea en el Perú a partir del repositorio en línea del ANA. Estos informes fueron publicados por el Instituto Nacional de Recursos Naturales (INRENA) contienen los resultados de ensayos efectuados entre los años 1997 a 2006 para evaluar el estado actual de las fuentes de agua subterránea en distintas regiones del Perú (Minagri, 2018). Un resumen de las características de los informes consultados se presenta en la tabla 4.1.

Tabla 4.1. Compilación de informes hidrogeológicos de la ANA recopilados para el presente estudio.

Zona de estudio	Fecha de publicación
Acarí	2003
Asia-Omas	2002
Casma	2003
Chancay	2004
Chao	1998
Chilca	2005
Chillón	2004
Chincha	2000
Huarmey	2002
Huaura	2005
Iquitos	2006
Jequetepeque	2005
La Leche	1999

Zona de estudio	Fecha de publicación
Lacramarca	2001
Lurín	2005
Mala	2002
Medio y bajo Piura	2004
Moche	2005
Palpa	2000
Pativilca	2005
Ramis	2004
Santa	2001
Supe	2005
Virú	1999
Yauca	2002

En el anexo de estos documentos, se presenta un listado de todos los pozos correspondientes a la zona de estudio de cada informe junto con sus características respectivas (año de perforación, nivel estático de agua, estado del pozo, caudal de explotación, entre otras). A partir de este compendio, se seleccionaron únicamente a los pozos con código de

Inventario de Recursos Hídricos subterráneos (IRHS), información de caudal y fecha de ejecución ensayos para la construcción de los modelos. En lo referido al caudal, al ser esta la variable dependiente en esta investigación, todo pozo sin este dato es inservible para el entrenamiento, validación y evaluación de los modelos de clasificación del potencial de agua subterránea. Respecto a la fecha de ejecución de ensayos, se descartaron a todos los pozos que carecían de este dato puesto que en la investigación se emplean valores independientes cuyo valor está condicionado al tiempo, como es el caso de la precipitación. Al concluir esta etapa, se obtuvo un total de 4476 pozos de baja profundidad en todo el Perú aptos para la construcción de los modelos de aprendizaje automatizado. En las figuras 4.2 y 4.3, se presenta la distribución de los 4476 de pozos de acuerdo con la zona de estudio y fecha de ejecución de estudios hidrológicos.

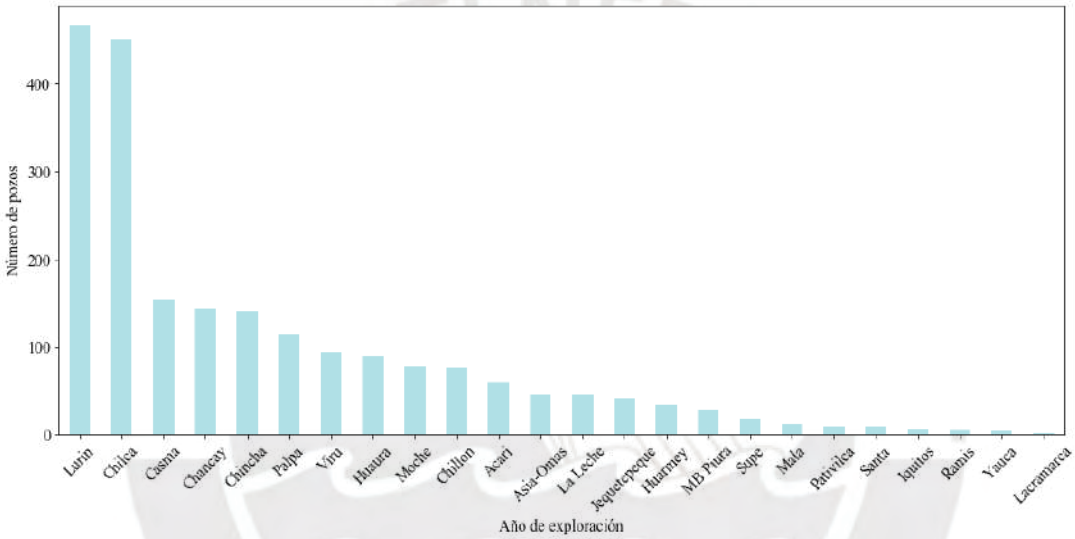


Figura 4.2. Distribución de pozos de acuerdo con la zona de estudio.

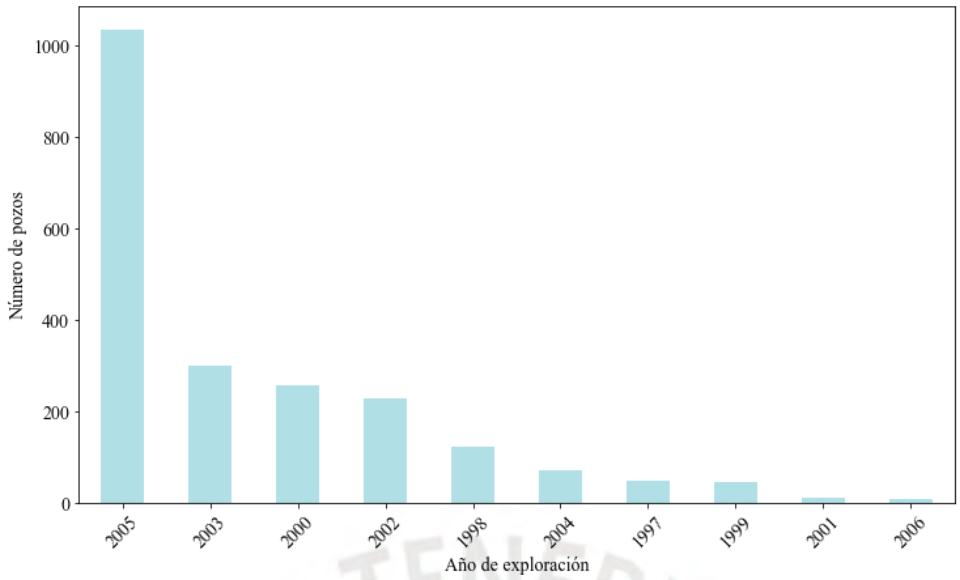


Figura 4.3. Distribución de pozos de acuerdo con el año que se ejecutaron los ensayos hidrológicos correspondientes.

Identificación de coordenadas de los pozos seleccionados

Puesto que las variables independientes que se emplearon en el modelo de clasificación dependen de la ubicación de los pozos, fue indispensable obtener sus coordenadas. Para ello, se recurrió al registro de pozos a nivel nacional publicado por la ANA (Castañeda y Céspedes, 2017). Esta base de datos contiene a los pozos de agua subterránea existentes en el Perú y detalla sus características tales como tipo de pozo, estado de explotación, propietario y código IRHS. Asimismo, se indican las coordenadas geográficas de cada uno en el sistema WGS84. Utilizando el código IRHS fue posible identificar las coordenadas para 3043 de los 4476 pozos seleccionados en la etapa anterior. En la figura 4.4, se presenta un mapa con la ubicación de los pozos. Claramente, se puede observar que la mayoría de estos se ubican en la costa del Perú. Asimismo, los pocos pozos correspondientes a las regiones de la sierra y selva se encuentran en los departamentos de Puno y Loreto respectivamente.

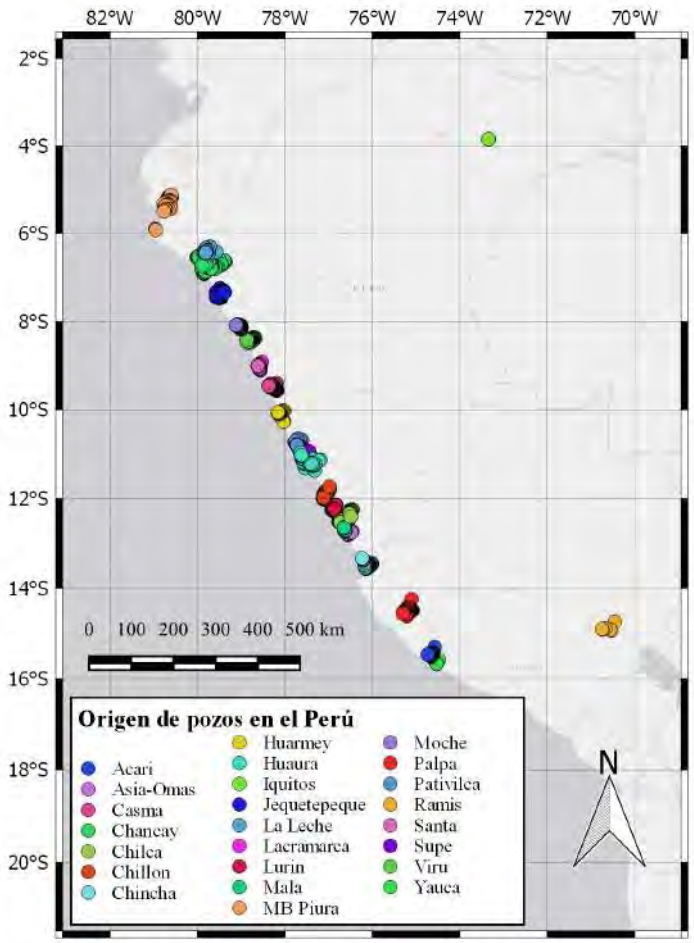


Figura 4.4. Origen de los pozos seleccionados para la construcción de los modelos de clasificación.

Categorización de los caudales recopilados

En esta investigación, los algoritmos de aprendizaje automatizado se emplearon para la aplicación de modelos de clasificación. Por este motivo, a nivel nacional, se optó por clasificar los caudales de los pozos seleccionados en tres categorías divididas uniformemente en base a los percentiles 33 y 66 de la variable caudal: categoría A (alto potencial de agua subterránea), categoría B (potencial medio) y categoría C (potencial bajo). En la figura 4.5, se presenta a modo de ejemplo la ubicación de cada pozo y su respectiva categoría en la localidad de Catacaos, Piura. Por otro lado, además de los modelos construidos con la totalidad de pozos del Perú, también se optó por elaborar modelos de clasificación especializados en regiones más pequeñas para analizar si la reducción del área de estudio permitía mejorar el desempeño de los modelos al disminuir la dispersión de las variables. Sin embargo, en el caso de los modelos planteados para la costa centro y costa norte, la escasez de pozos y la presencia de caudales repetidos provocó que la clase superior se encuentre subrepresentada en comparación a las otras dos. Por este motivo, se optó por limitar el alcance de la clasificación únicamente a

dos categorías: categoría A y categoría B para caracterizar a los emplazamientos con potencial de agua subterránea superior e inferior a la mediana respectivamente en estos casos.

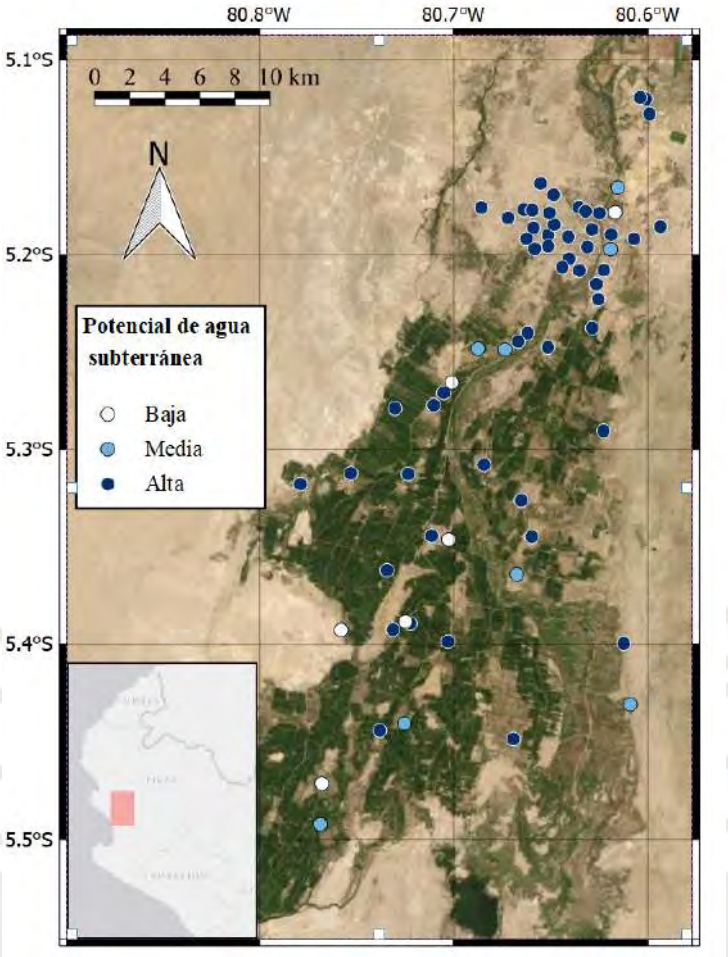


Figura 4.5. Ejemplo de categorización del potencial de agua subterránea de acuerdo con el caudal de los pozos para la ciudad de Catacaos en Piura.

Observaciones

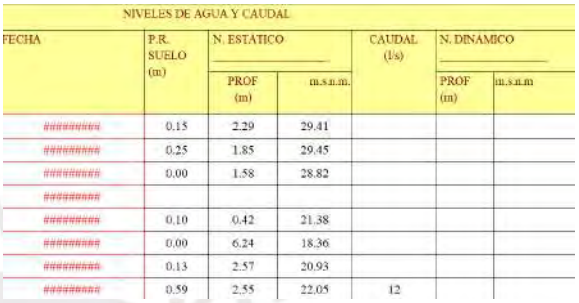
Concluida esta fase, se pudo apreciar que existe un manejo inadecuado de la información existente del agua subterránea en el Perú por parte de la ANA. En particular, de encontrarse toda la información adecuadamente registrada, se podrían emplear los 52 930 pozos almacenados de la base de datos nacional para la construcción de los modelos de clasificación. A continuación, se presentan las observaciones más resultantes:

- No se dispone de un documento que liste la totalidad de informes de pozos de agua subterránea en el Perú. Se desconoce si en la presente investigación se está trabajando con la totalidad de pozos con caudal conocido a nivel nacional.
- En lo que respecta a la calidad de la información presentada en cada informe, cabe resaltar que múltiples de estos no presentan un acabado adecuado. Algunos de los

problemas encontrados son los siguientes: valores de caudal reportados como letras o signos en vez de números, ausencia de anexos con información de caudal en múltiples informes, pozos sin fecha de estudio y caudales sumamente atípicos. Asesorando la calidad de la página web del "Observatorio de aguas subterráneas" donde se encuentran registradas las coordenadas de los pozos, se observó que algunos nombres de cuencas estaban incorrectamente redactados. Ejemplos de todas estas observaciones se presentan en la figura 4.6.

NIVELES DE AGUA Y CAUDAL				
N. ESTÁTICO		CAUDAL (l/s)	N. DINÁMICO	
PROF(m)	m.s.n.m		PROF(m)	m.s.n.m.
6.28		f		
4.60		5.00		
8.40				
6.60				
5.40				
5.52				
3.64				
3.20		6.00	4.80	

(a) Datos de caudal no numéricos



NIVELES DE AGUA Y CAUDAL						
FECHA	P.R. SUELO (m)	N. ESTÁTICO		CAUDAL (l/s)	N. DINÁMICO	
		PROF (m)	m.s.n.m.		PROF (m)	m.s.n.m
##########	0.15	2.29	29.41			
##########	0.25	1.85	29.45			
##########	0.00	1.58	28.82			
##########						
##########	0.10	0.42	21.38			
##########	0.00	6.24	18.36			
##########	0.13	2.57	20.93			
##########	0.59	2.55	22.05	12		

(b) Columna de fechas rellenada inadecuadamente

(c) Portal con datos incorrectos

Figura 4.6. Anomalías encontradas en los informes hidrogeológicos y en la base de datos de pozos de la ANA.

- Se desconoce el método de segmentación geográfica empleado para la creación de cada informe. Asumir que cada informe está asociado a cada uno de los acuíferos del Perú como aparentemente indica la tendencia no es del todo correcto dado que existen informes que no corresponden a ninguno de los acuíferos identificados en el "Observatorio de aguas subterráneas" de la ANA. Estos son los informes de Caravelí, Chao, Piura, Pucalpa, Coata y Lanchas. Además, existen archivos incorrectamente clasificados en el repositorio en línea de la ANA. Por ejemplo, el vínculo que debería descargar el documento titulado como "Inventario de fuentes de agua subterránea en el valle Piura (parte alta): informe final" en realidad descarga el informe correspondiente al valle de Acarí.
- Finalmente, cabe resaltar que no se pudieron emplear los 4476 pozos con caudal

registrado en los informes dado que su código IRHS no se encontraba en la base de datos del "Observatorio de aguas subterráneas". En consecuencia, no se disponen las coordenadas de estos pozos, las cuales son imperativas para la asignación de variables independientes en la siguiente sección. Los pozos cuya información no fue aprovechable debido a la ausencia de coordenadas fueron 1433: 37 para Acarí, 671 para Chao, 148 para Cañete, 5 para Chillón, 6 para Chincha, 5 para Huaura, 41 para Loreto, 15 para Lacramarca, 139 para Lurín, 9 para Medio y Bajo Piura. 26 para Moche, 12 para Palpa, 105 para Santa, 2 para Ramis, 1 para Jequetepeque y 211 para La Leche.

4.1.2. Recopilación de las variables independientes

En esta etapa, se recopilaron todas las propiedades correspondientes a la ubicación de los pozos que estuviesen relacionadas con el potencial del agua subterránea. Estas propiedades pueden ser categorizadas en su mayoría como topográficas, hidrológicas, geológicas, pedológicas y ambientales. La obtención de estos parámetros se realizó en base a fuentes públicas de acceso gratuito de información georreferenciada. Un resumen de todas estas fuentes se presenta en la tabla 4.2. Posteriormente, estos datos fueron procesados en la plataforma del Sistema de Información Geográfica (SIG) de *Google Earth Engine* (GEE) (Gorelick et al., 2017) para la construcción de la base de datos empleada para el desarrollo de los modelos de clasificación. El procesamiento de las variables independientes se realizó utilizando el mismo sistema de coordenadas empleado por el ANA para la georreferenciación de pozos: WGS 1984, también conocido como EPSG:4326. Asimismo, se empleó QGIS (QGIS Development Team, 2020) para el diseño de los mapas presentados en la investigación. El código empleado en esta etapa fue desarrollado en JavaScript y puede ser accedido en el siguiente enlace https://code.earthengine.google.com/fe63cd6184b009824ed3c843fdc5544d.

Tabla 4.2. Metadatos de la información georreferenciada utilizada.

Nombre	Formato	Fuente	Resolución	Fecha
Acumulación de flujo	Ráster	Lehner et al. (2008)	500 m	2000
Contenido de arcilla en el suelo	Ráster	Hengl (2018a)	250 m	1950-2018
Contenido de arena en el suelo	Ráster	Hengl (2018b)	250 m	1950-2018
Elevación digital SRTM	Ráster	Farr et al. (2007)	30 m	2000
Evapotranspiración neta	Ráster	Running et al. (2017)	500 m	2001-2020
Geomorfología SRTM	Ráster	Theobald et al. (2015)	90 m	2006-2011
NDVI	Ráster	Didan (2015)	250 m	2000-2020
NDWI	Ráster	Chander et al. (2009)	15 m	2013-2018
Precipitación mensual	Ráster	C3S (2017)	25 km	1979-2020
Textura del suelo USDA	Ráster	Hengl (2018c)	250 m	1950-2018
Fallas geológicas del Perú	Vector	INGEMMET (2019)	No aplica	2017
Grupos estratigráficos del Perú	Vector	INGEMMET (2019)	No aplica	2017
Humedad del suelo	Vector	Hengl y Gupta (2019)	250 m	1950-2018

A partir de las bases de datos consultadas, se obtuvieron las variables independientes presentadas en la tabla 4.3. Estas se clasifican dos tipos: continuas o categóricas. En el caso de las variables continuas, cada píxel contenido en el territorio nacional posee un valor numérico que cuantifica a la característica correspondiente en un lugar determinado. Por ejemplo, elevación es una variable continua, razón por la cual un píxel a nivel del mar en la ciudad de Lima tendría un valor de 0, representando una elevación de 0 metros Por otro lado, al tratar con variables categóricas, el número correspondiente a cada píxel corresponde a una clase definida por lo que los valores que este puede adoptar son limitados. Por ejemplo, en el caso de la variable tipo de suelo, un valor de 0 representa un suelo arcilloso en el píxel. Cabe resaltar que todas las variables empleadas en la presente investigación fueron seleccionadas en base a las empleadas para la aplicación de modelos satisfactorios de clasificación de potencial de agua subterránea en otros países (ver tabla 3.1), tomando en cuenta la disponibilidad de datos para el Perú.

Tabla 4.3. Clasificación de variables de acuerdo con el tipo de datos.

Variables topográficas		**Variables hidrológicas**		**Variables geológicas y otras**	
Nombre	**Tipo**	**Nombre**	**Tipo**	**Nombre**	**Tipo**
Elevación	Continua	Densidad de drenaje	Continua	Distancia a fallas	Continua
Pendiente	Continua	Distancia a ríos	Continua	Tipo de suelo	Categórica
Aspecto	Continua	Evapotranspiración	Continua	Contenido de arena	Continua
Tipo de relieve	Categórica	Índice diferencial de agua normalizado	Continua	Contenido de arcilla	Continua
		Precipitación	Continua	Formación geológica	Categórica
		Índice de fuerza de corriente	Continua	Humedad del suelo	Continua
		Índice topográfico de humedad	Continua		

A continuación, se presenta la relación de cada variable independiente empleada con el potencial de agua subterránea y la distribución espacial de las variables en el territorio nacional. En el caso de las variables continuas, los mapas presentados en las páginas siguientes corresponden a una visualización con límites inferior y superior correspondientes a los cuartiles 1 y 3 respectivamente. En el caso de las variables categóricas, cada color corresponde a una clasificación distinta de acuerdo con la leyenda presentada.

Variables topográficas (figura 4.8)

- **(4.8a) Elevación:** Entiéndase como la distancia vertical de un objeto medido desde el nivel del mar. Esta propiedad posee una relación indirecta con el potencial de agua subterránea de un sector. A mayor altura, la reducción de la densidad de vegetación y la acumulación de precipitación y nieve favorece la infiltración de agua a los acuíferos (Liniger y Weintgartner, 2008). Para los propósitos de esta investigación, se empleó la información correspondiente del *Shuttle Radar Topography Mission* (SRTM) de la NASA.

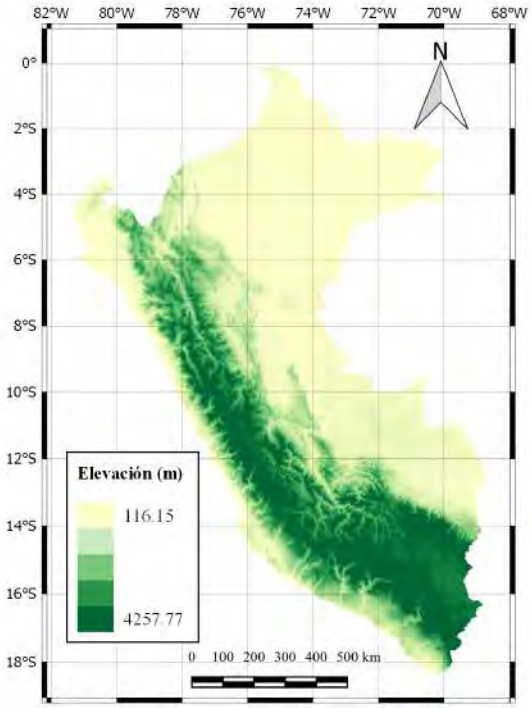
- **(4.8b) Pendiente:** La relación entre la pendiente de la cuenca superficial y el potencial de agua subterránea es sumamente compleja. En suelos permeables, la capacidad de infiltración disminuye con el aumento de la pendiente hasta un cierto valor límite que depende del tipo de suelo. En cambio, para suelos impermeables, se observa una relación muy limitada entre la pendiente y la capacidad de infiltración del agua superficial a acuíferos (Nassif y Wilson, 1975). En este proyecto, se determinó la

pendiente calculando la primera derivada de la información de elevación SRTM tomando en cuenta cuatro píxeles adyacentes.

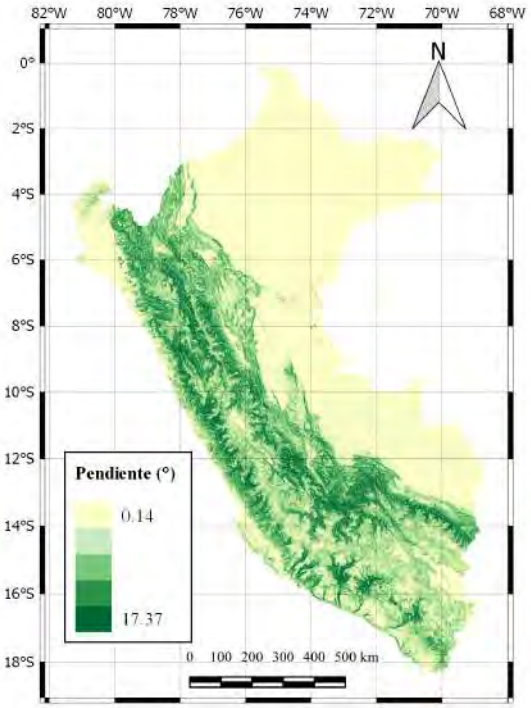
- **(4.8c) Aspecto:** Entiéndase como la orientación de la recta de máxima pendiente en relación con el norte geográfico. Se ha observado que el aspecto posee una íntima relación con los microclimas de una región. En el hemisferio sur, las pendientes con aspecto orientado al norte tienden a encontrarse más expuestas a la radiación del sol y a desarrollar mayores niveles de evapotranspiración en comparación a las pendientes de orientación sur (Bennie, Hill, Baxter, y Huntley, 2006). En consecuencia, se puede esperar un menor potencial de agua subterránea mientras el aspecto tienda más al valor de cero grados sexagesimales.

- **(4.8d) Tipo de relieve:** Propiedad cualitativa definida de acuerdo con la geomorfología del terreno. En base a la información topográfica disponible, se clasifica cada sector como montaña, acantilado, valle, cima, etc. Puesto que la geomorfología del terreno determina la facilidad con la que el agua se acumula en un lugar determinado, esta propiedad actúa como indicador de percolación del agua hacia el subsuelo (McLachlan y Brown, 2006). En la presente investigación, se utilizó la clasificación realizada por la organización *Conservation Science Partners* (CSP) en base a los datos SRTM. En esta, se clasifica el relieve en 15 categorías distintas según el índice topográfico de posición (TPI) y el índice continuo de calentamiento/aislamiento (CHILI). Para ejemplificar de mejor manera esta categorización, se presenta en la figura 4.7 la ciudad de Lurín clasificada de acuerdo con el tipo de relieve.

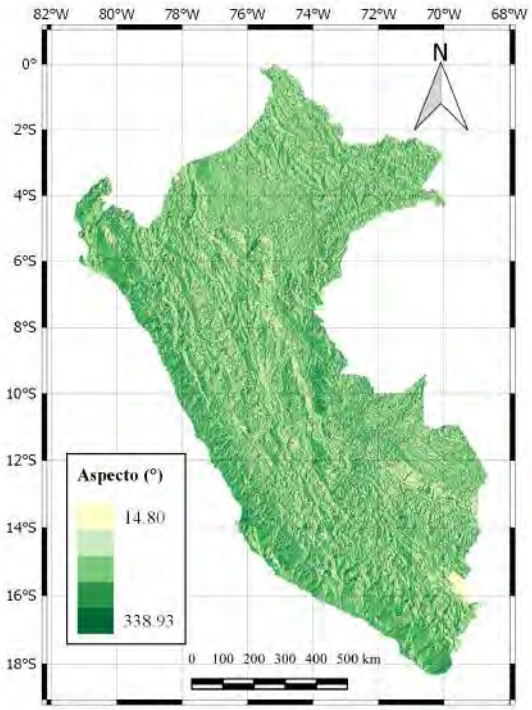
Figura 4.7. Clasificación por relieve de la ciudad de Lurín.



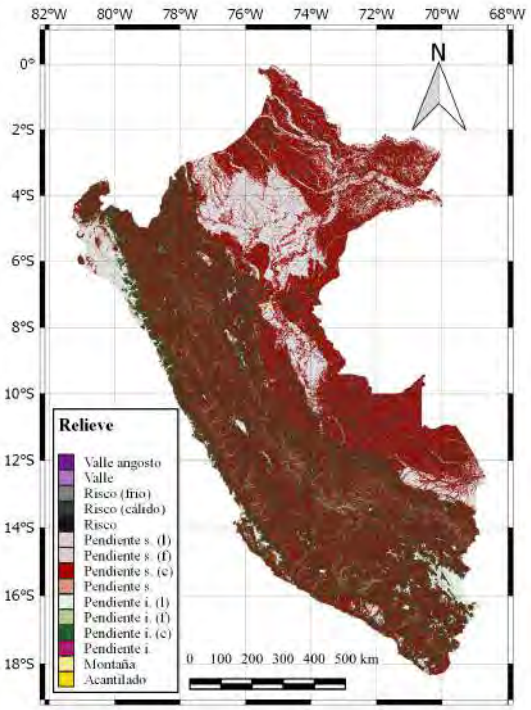
(a) Elevación en metros



(b) Pendiente en grados sexagesimales



(c) Aspecto en grados sexagesimales



(d) Tipo de relieve

Figura 4.8. Variables topográficas empleadas para el entrenamiento de los modelos.

Variables hidrológicas (figura 4.9)

- **(4.9a) Densidad de drenaje:** Parámetro definido a través de la ecuación 4.1 por Horton (1945):

$$D_d = \frac{\sum_i l_i}{A} \tag{4.1}$$

  Donde $\sum_i l_i$ es la suma de la longitud de todos los cauces comprendidos en la cuenca hidrográfica y A es el área de esta. Esta característica indica el grado de eficiencia de la red de drenaje en una cuenca, por lo que un alto valor de densidad de drenaje supone que una mayor parte de la precipitación se convierta en escorrentía y no contribuya a la recarga de los acuíferos. En consecuencia, secciones bien drenadas tienden a tener un potencial de agua subterránea más limitado (Kakish y Katimbo, 2017). En la presente investigación se calcularon los valores de densidad de drenaje empleando la información espacial de ríos y subcuencas del Instituto Geográfico Nacional (IGN).

- **(4.9b) Distancia a ríos:** Debido a la conexión entre el agua superficial y el agua subterránea a través del ciclo hidrológico, las áreas más cercanas a cuerpos de agua dulce tienden a disponer de un volumen de infiltración potencial mayor. Por ende, conforman sectores más adecuados para la construcción de pozos. El agua percolada durante la época de crecidas permite que aún en época de estiaje sea posible disponer de un caudal de explotación adecuado en estos lugares (Dijon, 1981). Para el cálculo de este parámetro, se transformaron los vectores de ríos del IGN a píxeles y se creó un ráster cuyo valor de celda corresponde a la distancia euclidiana entre cada píxel hacia el río más cercano.

- **(4.9c) Evapotranspiración:** Proceso definido por Chow (1988) como la combinación de procesos de evaporación de la superficie del suelo y transpiración vegetal. Elevadas tasas de evapotranspiración suelen asociarse con un menor potencial de agua subterránea puesto que reduce la tasa de recarga del acuífero. No obstante, se ha demostrado que la sensibilidad de este último parámetro frente a la evapotranspiración depende en gran medida del tipo de suelo (Hartmann, Gleeson, Wada, y Wagener, 2017). Por ejemplo, suelos granulares de rápida saturación suelen facilitar la percolación del agua aún frente a eventos de precipitación moderadamente baja. La tasa de evapotranspiración empleada en la presente investigación se expresa en kg/m^2. Los valores computados corresponden al resultado de aplicación de la ecuación de Penman-Monteith en base a la información geoespacial recibida por MODIS desde el año 2001 hasta el presente. La expresión matemática correspondiente al cálculo de la evapotranspiración se presenta en la ecuación 4.2.

$$ET_0 = \frac{\Delta(R_n - G) + \rho c_p(\delta e) g_a}{(\Delta + \gamma(1 + g_a/g_s))L_v} \tag{4.2}$$

Donde Δ es la tasa de cambio de la humedad específica de saturación, R_n es la radiación neta, G es el flujo de calor, c_p es la capacidad calorífica específica del aire, p_a es la densidad seca del aire, δe es el déficit de presión de vapor, g_a es la conductividad del aire, g_s es la conductividad de los estomas y γ es la constante psicométrica. El ráster empleado para el modelo de clasificación se computó como el promedio histórico hasta la fecha para todo el territorio nacional.

- **(4.9d) Índice diferencial de agua normalizado:** Más conocido como NDWI por sus siglas en inglés, este parámetro fue definido por McFeeters (1996) para la ubicación de áreas con alto potencial de inundación basándose únicamente en información satelital. El valor de NDWI se define en la ecuación 4.3.

$$NDWI = \frac{v - NIR}{v + NIR} \tag{4.3}$$

Donde v es la banda verde y NIR corresponde a la banda de infrarrojo cercano del satélite, razón por las cuales estos valores oscilan entre -1 (bajo potencial de inundación) a 1 (alto potencial de inundación). Se espera que en emplazamientos con mayores niveles de NDWI se obtenga un mayor potencial de agua subterránea gracias al empozamiento que favorece la percolación. Hartmann et al. (2017) emplearon este parámetro satisfactoriamente para la estimación del nivel piezométrico de acuíferos en Punjab, India.

- **(4.9e) Precipitación acumulada mensual:** Entiéndase como una medida para cuantificar la caída acumulada del agua en cualquier estado físico en un área geográfica durante un intervalo de tiempo fijo de un mes (Chow, 1988). Dependiendo de las características geológicas y pedológicas del emplazamiento, la precipitación puede conformar hasta el 40 % de la fuente total de recarga de un acuífero (Kotchoni et al., 2019). Puesto que esta investigación buscó emplear únicamente variables telemedidas, la información de precipitación para la presente investigación fue extraída de la base de datos ERA5 del Servicio de Cambio Climático COPERNICUS de la Unión Europea. Esta computa sus resultados en base a la información de múltiples agencias satelitales que incluyen a NASA, JAXA, NOAA, EUMETSAT, entre otros. Para ello, aproxima la precipitación empleando modelos numéricos en base las propiedades atmosféricas de presión, temperatura y humedad y las ecuaciones propuestas por el Sistema Integrado de Predicción Meteorológica (IFS) (Hennermann y Guillory, 2020). Al tratarse de una variable temporal importante, se emplearon cuatro variaciones de la variable precipitación para evaluar su influencia en el desempeño del modelo:

1. Valor promedio de precipitación histórica hasta la actualidad
2. Valor exacto de precipitación correspondiente al mes en el cual se midió el caudal de explotación de los pozos
3. Promedio anual correspondiente al año en el cual se midió el caudal de explotación de los pozos
4. Promedio del mes en el cual se midió el caudal de explotación de los pozos en un rango de +/- 5 años

- **(4.9f) Índice de fuerza de corriente:** Más conocido como SPI por sus siglas en inglés (*Stream Power Index*), este parámetro se define en la ecuación 4.4:

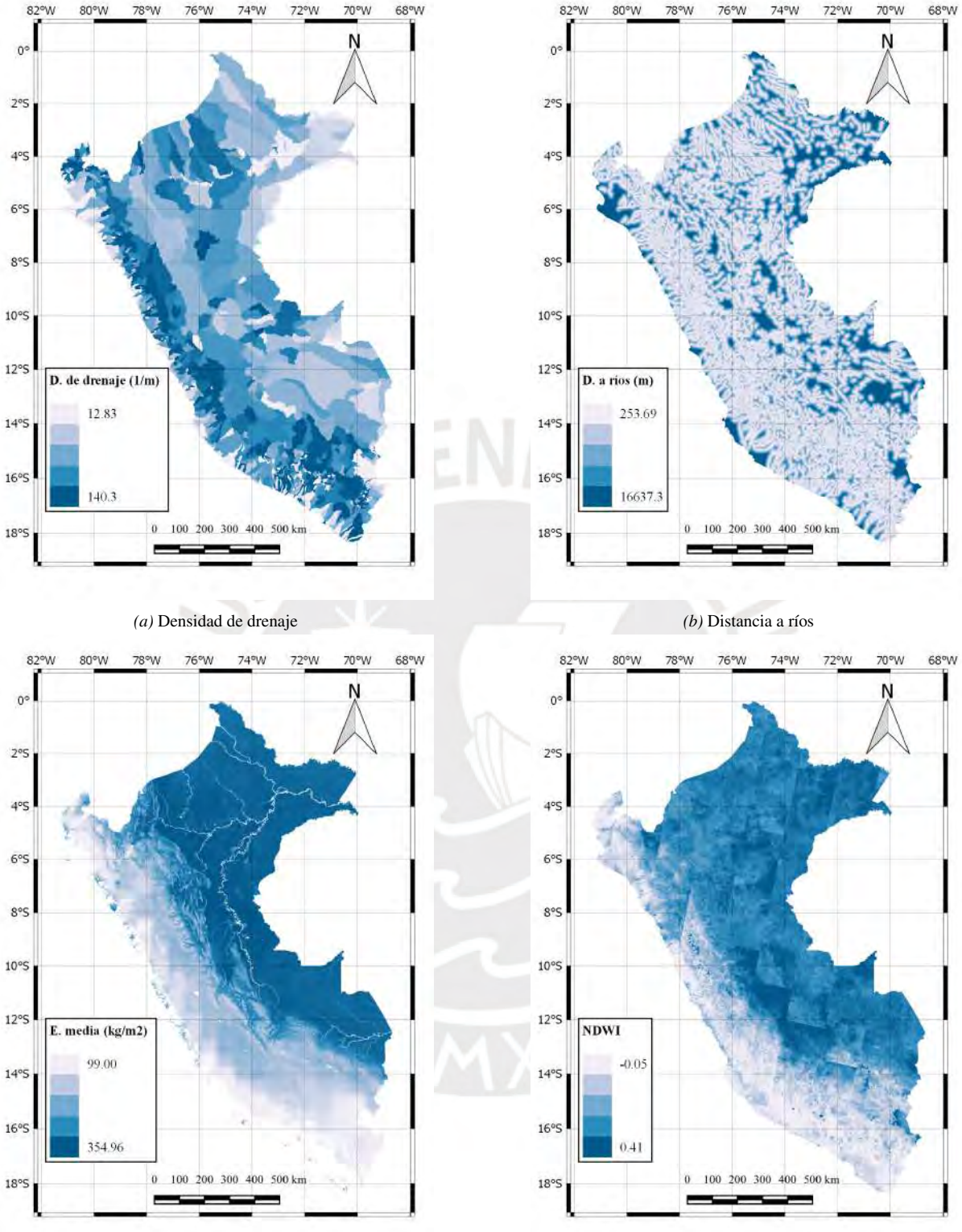
$$SPI = UCA.tan(b) \tag{4.4}$$

Donde UCA es el área de contribución de drenaje áreas arriba y b es la pendiente. El SPI se entiende como una medida del poder erosivo del agua que fluye y aumenta conforme es mayor tanto el drenaje aguas arriba como la pendiente del punto de evaluación (Bustos y Bermúdez, 2017). Las altas velocidades de zonas con alto SPI suponen tasas bajas de infiltración a los acuíferos, afectando negativamente al potencial de agua subterránea. Para los propósitos de la presente investigación, se optó por utilizar el logaritmo natural de este parámetro ya que los valores obtenidos oscilaban en un rango muy amplio que imposibilitaba su visualización adecuada en el territorio peruano.

- **(4.9g) Índice topográfico de humedad:** Abreviado como TWI por sus siglas en inglés (*Topographic Wetness Index)*, este parámetro se define en la ecuación 4.5:

$$TWI = ln(\frac{UCA}{tan(b)}) \tag{4.5}$$

Donde nuevamente UCA es el área de contribución de drenaje áreas arriba y b es la pendiente (Rahmati et al., 2016). Altos valores de TWI se pueden evidenciar en las zonas con mayor susceptibilidad a ser inundadas. En otras palabras, estas zonas facilitan la acumulación y percolación del agua superficial, lo cual favorece a su potencial de agua subterránea.



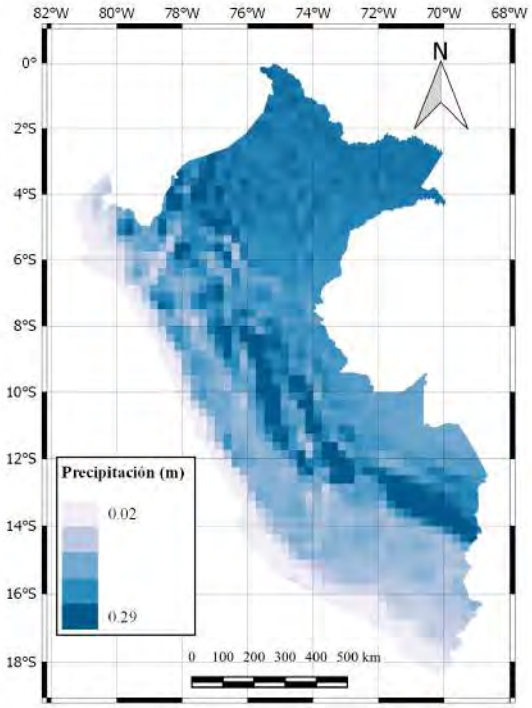
(a) Densidad de drenaje

(b) Distancia a ríos

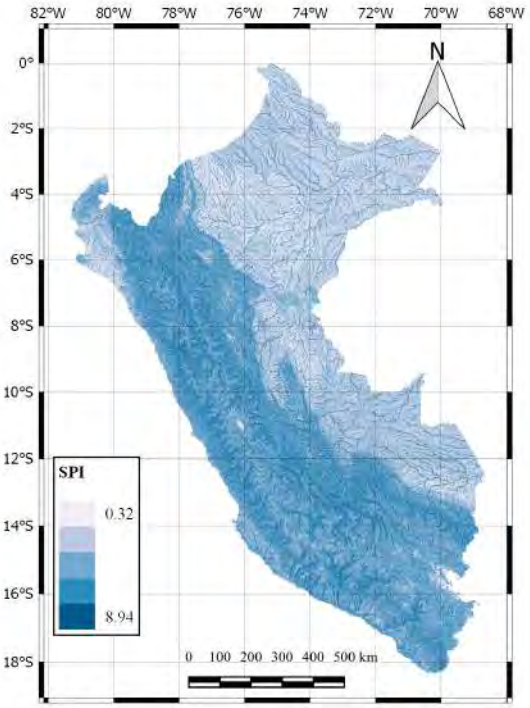
(c) Evapotranspiración media

(d) NDWI

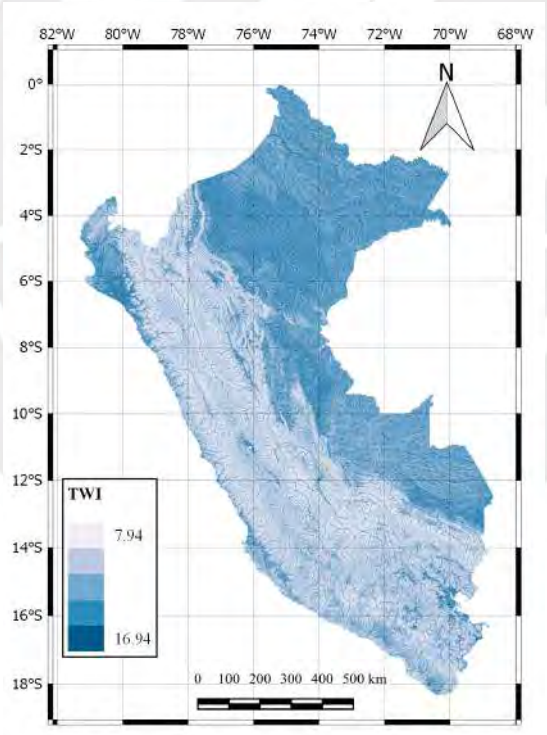
Figura 4.9. Variables hidrológicas empleadas para el entrenamiento de los modelos.



(e) Precipitación mensual acumulada



(f) SPI



(g) TWI

Figura 4.9 (cont.). Variables hidrológicas empleadas para el entrenamiento de los modelos.

Variables geológicas, pedológicas y ambientales (figura 4.11)

- **(4.11a) Distancia a fallas:** Una falla geológica se define como una fractura en la corteza terrestre debido al movimiento de placas tectónicas (USGS, 2019). Cuando esta deformación se produce en la corteza superficial (profundidad menor a 1 km), se introduce permeabilidad, heterogeneidad y anisotropía a la roca afectada. En consecuencia, se facilita la conducción del agua a través de las formaciones rocosas a lo largo de la falla. Sin embargo, en el sentido perpendicular se produce el efecto contrario, el desplazamiento de la corteza forma barreras que limitan el flujo subterráneo (Bense, Gleeson, Loveless, Bour, y Scibek, 2013). En la presente investigación, se obtuvo los datos de ubicación de las fallas geológicas del Perú a través del Sistema de Información Geológico y Catastral Minero (GEOCATMIN), este incluye fallas normales, inversas y de rumbo (INGEMMET, 2019). Luego, se determinó la distancia euclidiana en metros entre cada píxel y la falla más cercana.

- **(4.11b) Tipo de suelo según textura:** Propiedad relacionada a la distribución de las partículas inorgánicas del suelo; es decir, grava, arena, arcilla y limo. El Departamento de Agricultura de los Estados Unidos (USDA) propone una clasificación basada en 12 clases de textura de acuerdo con el porcentaje de arena, arcilla y limo. En la figura 4.10 se presenta cómo funciona este sistema de clasificación.

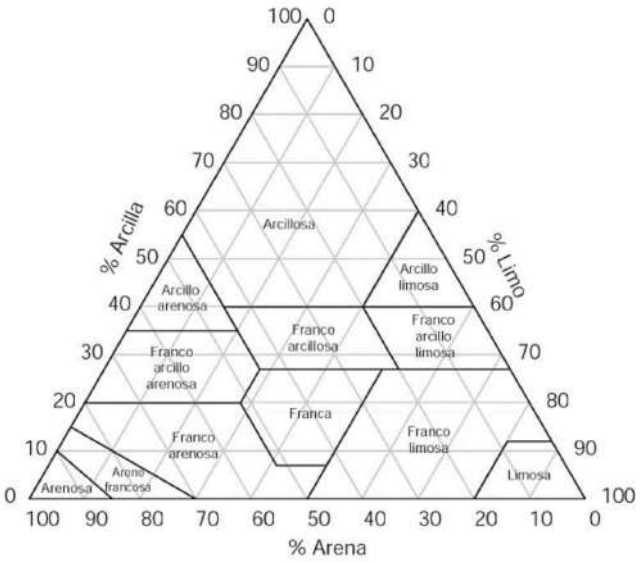


Figura 4.10. Pirámide de clasificación del tipo de suelo según el USDA

Esta propiedad se relaciona íntimamente con la medida de permeabilidad del suelo, la cual se representa por la constante de Darcy presente en la ecuación 4.6:

$$U = -K\frac{dh}{dz} \tag{4.6}$$

Donde U es la velocidad del flujo subterráneo, K es la conductividad hidráulica, h es la

carga hidráulica y z es la distancia vertical al suelo. Si bien el valor exacto de K debe ser determinado a través de un ensayo de permeabilidad en un laboratorio de mecánica de suelos, este valor depende íntimamente del tipo de suelo. Por ejemplo, al disponer de partículas muy porosas, los suelos granulares se asocian a un mayor valor de K en comparación a los suelos cohesivos. Por ello, los suelos arenosos se caracterizan por tener mejor conducción de fluidos y favorecer el potencial de agua subterránea (Spiegel, 2017). Para la presente investigación, se empleó un ráster con resolución espacial de 250 m desarrollado por EnvirometriX Ltd en base a la clasificación del USDA para una profundidad de 200m (Hengl, 2018c). Los códigos asociados a cada tipo de suelo que fueron empleados en la leyenda del mapa presentado en la figura 4.11b corresponden a abreviaciones cuyos significados se presentan en la tabla 4.4.

Tabla 4.4. Interpretación de los códigos de tipo de suelo empleados en la figura 4.11b

Código	**Clasificación en inglés**	**Clasificación en español**
Cl	Clay	Arcilloso
SiCl	Silty clay	Arcillo limoso
SaCl	Sandy clay	Arcillo arenoso
ClLo	Clay loam	Franco arcilloso
SiClLo	Silty clay loam	Franco arcilloso limoso
SaClLo	Sandy clay loam	Franco arcillo arenoso
Lo	Loam	Franca
SiLo	Silt loam	Franco limoso
SaLo	Sandy loam	Franco arenoso
Si	Silt	Limoso
LoSa	Loamy sand	Arenoso franco
Sa	Sand	Arenoso

- **(4.11c) Contenido porcentual de arena:** Propiedad obtenida por Hengl (2018b) mediante modelos de aprendizaje automatizado en base a una compilación global de perfiles y muestras de suelo para una profundidad de 200m. Al disponer de partículas de mayor tamaño que las arcillas, se espera que su alta porosidad facilite la conducción de agua, significando ello a su vez en un mayor potencial de agua subterránea.
- **(4.11d) Contenido porcentual de arcilla:** Propiedad obtenida también por Hengl (2018a) mediante modelos de aprendizaje automatizado en base a una compilación global de perfiles y muestras de suelo para una profundidad de 200m. En contraste al caso del contenido de arena, se anticipa que suelos con una mayor cantidad de material

arcilloso dificulten el paso del agua a través de su estructura, afectando negativamente al caudal explotable de los acuíferos.

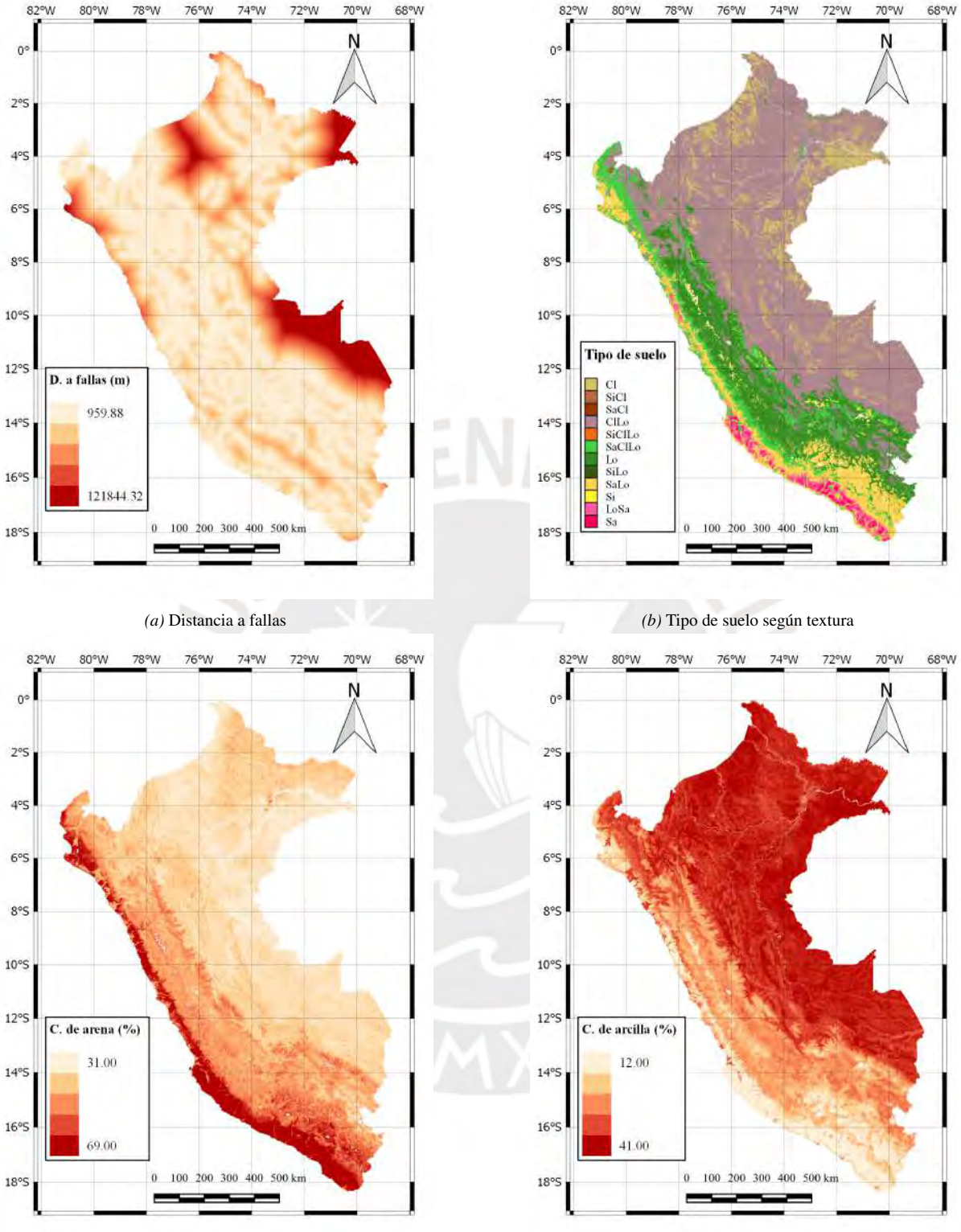
- **(4.11e) Contenido de humedad del suelo:** El contenido volumétrico porcentual estimado de agua en el suelo hasta una profundidad de 200m se aproximó por Hengl y Gupta (2019) en base a estudios de clasificación de suelos realizados por USDA, el Centro Internacional de Referencia e Información de Suelos (ISRIC, por sus siglas en inglés), entre otros. Se espera que, en áreas con mayor contenido volumétrico de agua, se favorezca el caudal de explotación de los acuíferos existentes.

- **(4.11f) Índice diferencial de vegetación normalizado:** Más conocido como NDVI por sus siglas en inglés (*Normalized Difference Vegetation Index*), este valor se define como un indicador de la cantidad vegetación presente en una determinada región. El cálculo se basa en el hecho de que la estructura de las hojas refleja predominantemente a las ondas de la región espectral del infrarrojo cercano (*NIR*) (Robinson et al., 2017). A este valor se le resta el de las ondas de color rojo (r) del espectro visible ya que estas son absorbidas por la vegetación. Finalmente, el parámetro NDVI se obtiene al dividir por la suma de ambos espectros, obteniendo la ecuación 4.7:

$$NDVI = \frac{NIR - r}{NIR + r} \tag{4.7}$$

  En un estudio realizado en los acuíferos de India, se demostró la existencia de una correlación íntima entre el valor del NDVI con los caudales de explotación de agua subterránea en 15000 pozos (Bhanja et al., 2019). Puesto que la presencia de agua subterránea facilita el crecimiento de la vegetación local, se espera que a mayor valor de NDVI, se obtenga un mayor potencial de agua subterránea. En la figura 4.11f se presenta el índice NDVI para el territorio nacional en base a la información del espectrorradiómetro de imágenes de media resolución (MODIS) de la NASA con una resolución de 250m. Análogamente al NDWI, este parámetro también oscila entre valores de -1 y +1.

- **(4.11g) Formaciones geológicas:** Entiéndase como la unidad mínima lito estratigráfica. Cada formación se define por un conjunto de estratos rocosos que poseen características físicas, químicas y biológicas similares (Parker, 1984). El movimiento del agua subterránea queda definido por las estructuras geológicas subterráneas de cada formación. Dependiendo de sus propiedades, estas pueden favorecer el flujo subterráneo o interrumpirlo en su totalidad (Folch, s.f.). Para la presente investigación, se empleó la información disponible en la plataforma GEOCATMIN para obtener la información geológica del territorio nacional (INGEMMET, 2019). En el anexo del presente documento, se presenta una tabla con los nombres completos de cada formación.



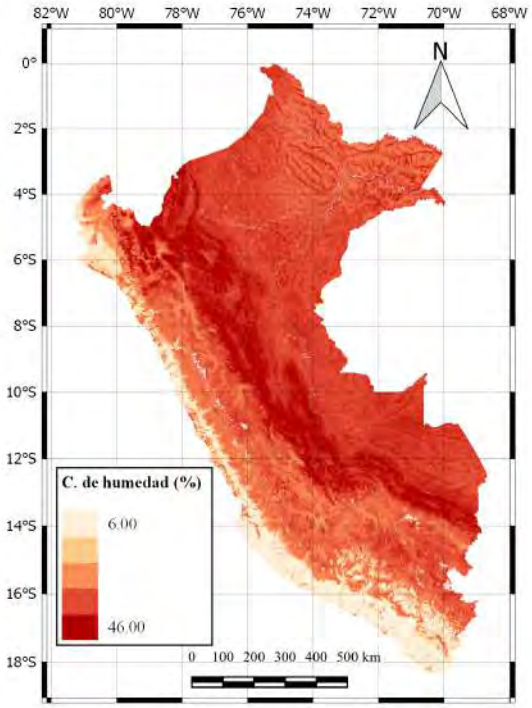
(a) Distancia a fallas

(b) Tipo de suelo según textura

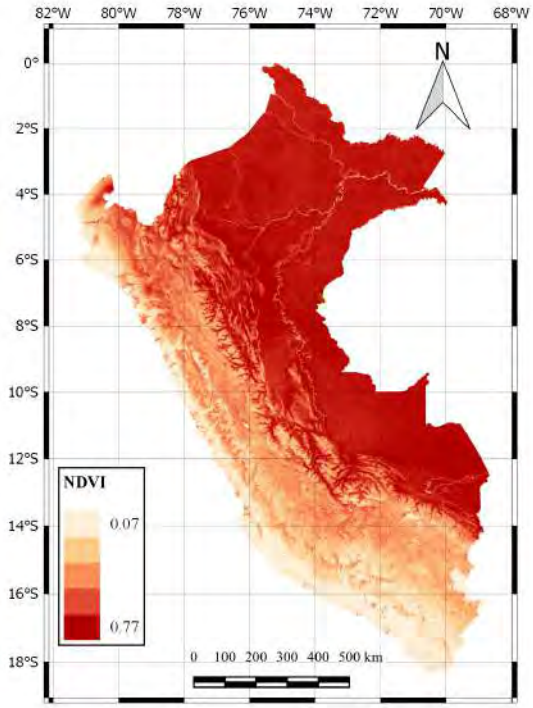
(c) Contenido porcentual de arena

(d) Contenido porcentual de arcilla

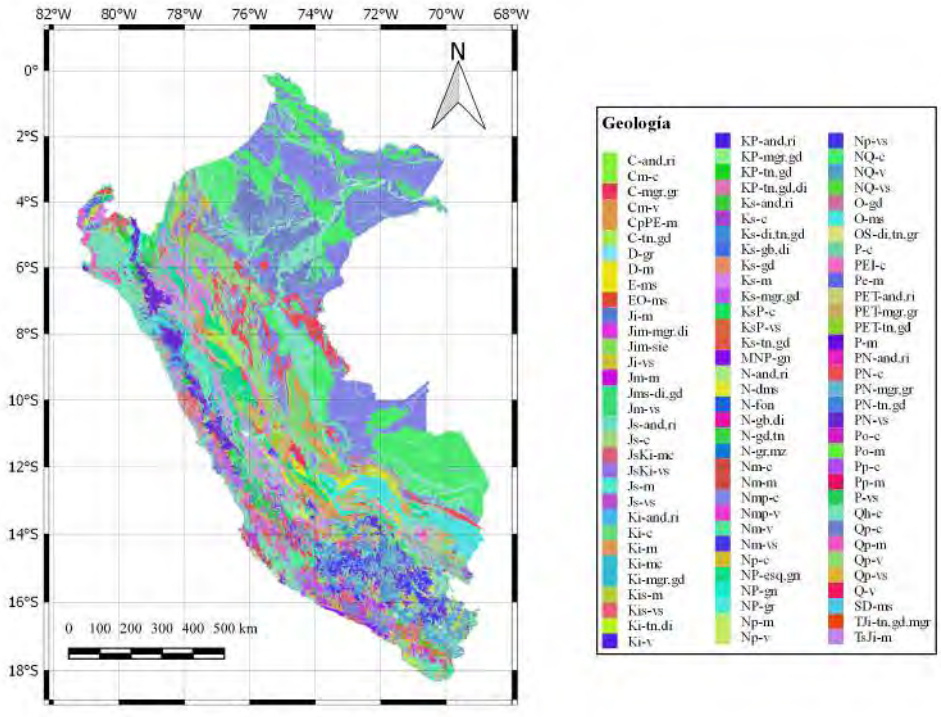
Figura 4.11. Variables geológicas, pedológicas y ambientales empleadas para el entrenamiento de los modelos.



(e) Contenido de humedad del suelo



(f) NDVI



(g) Formaciones geológicas

Figura 4.11 (cont.). Variables geológicas, pedológicas y ambientales empleadas para el entrenamiento de los modelos.

Finalizado este proceso, se observó que los píxeles de múltiples variables no cubrían de manera perfecta el territorio nacional. La presencia de huecos en múltiples sectores se explica debido a limitaciones existentes en la toma satelital de datos. Por este motivo, todos los pozos ubicados en emplazamientos con píxeles con valores nulos fueron descartados de la operación. En consecuencia, el número de pozos utilizables para la construcción de los modelos de clasificación se redujo a 2127.

4.1.3. Preprocesamiento de datos

Análisis estadístico de variables

Para explorar el comportamiento de las variables numéricas independientes y dependiente, se procedió a determinar los mínimos, máximos, promedios, desviaciones estándar, coeficientes de variación y cuartiles correspondientes. De esta manera, se cuantificó los órdenes de magnitud y la dispersión de estas. Por otro lado, para el análisis de las variables categóricas, se procedió a construir la distribución de frecuencias con el objetivo de conocer en qué categorías se concentran los pozos empleados en la construcción de los modelos.

Análisis de independencia de variables numéricas

Se analizó la presencia del problema de multicolinealidad entre las variables independientes de la base de datos utilizando dos procedimientos: una matriz de correlación de Pearson y el factor de inflación de la varianza (VIF, por sus siglas en inglés). A pesar de que el desempeño del modelo resultante no es afectado por la presencia de variable correlacionadas, sí altera la interpretabilidad de la importancia de cada variable independiente (Allen, 2007). Asimismo, el incluir dos variables con una alta correlación supone un incremento innecesario de la complejidad computacional del modelo y dificulta su entrenamiento. Si bien la matriz de correlación es el procedimiento tradicional para este análisis, este presenta dos limitaciones importantes. En primer lugar, no existe consenso acerca de un valor límite a partir del cual uno pueda afirmar que la correlación es suficientemente fuerte para ser perjudicial (Schober, Boer, y Schwarte, 2018). La segunda limitación se relaciona con el tipo de análisis, ya que en la matriz de correlación solo se determina si entre cada combinación de dos variables existe una correlación, mas no toma en cuenta combinaciones del resto de las variables en simultáneo. Para superar estas limitaciones, se determinó el parámetro VIF para cada variable independiente. En el cálculo del VIF, se examina en qué medida una determinada variable independiente puede ser construida a partir de una combinación lineal del resto de variables (Craney y Surles, 2002). Asimismo, la bibliografía indica que, para una determinada variable, valores de VIF que oscilan entre 5-10

evidencian una correlación moderada con el resto de variables, mientras que valores superiores a 10 indican una correlación fuerte. La formulación matemática de este parámetro se define en la ecuación 4.8.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{4.8}$$

Donde R_i^2 es el coeficiente de determinación correspondiente a una regresión lineal realizada con la variable iésima como dependiente y las sobrantes como independientes.

Análisis preliminar de importancia

Una estimación de la relevancia de las variables independientes continuas para la predicción de la variable dependiente se realizó con el objetivo de eliminar variables innecesarias que no contribuyan a la predicción, previamente a la construcción de modelos. Para ello, se recurrió al algoritmo Relief-F, puesto que permite computar un puntaje cuyo valor indica la calidad de cada una de las variables con relación a las demás para problemas de clasificación multiclase como es el presente (Kononenko, 1994). El funcionamiento de este algoritmo consiste en definir un vector de importancia de variables independientes (W_i) mediante un proceso iterativo. En cada iteración, se selecciona de manera aleatoria un elemento de la base de datos (x_i). Respecto a dicho elemento, se calcula la distancia euclidiana hacia todos los demás datos considerando todas las variables independientes. Concluido este proceso, se definen dos elementos denominados *coincidencia cercana* (c_i) y *discrepancia cercana* (d_i). La *coincidencia cercana* corresponde al elemento con menor distancia euclidiana y que posee la misma categoría que el dato seleccionada al azar. En cambio, la *discrepancia cercana* es el elemento que presenta la menor distancia euclidiana y que al mismo tiempo posee una categoría distinta. Luego, el peso de cada variable se actualiza de acuerdo con la ecuación 4.9.

$$W_i = W_i - (x_i - c_i)^2 + (x_i - d_i)^2 \tag{4.9}$$

De esta manera, el algoritmo beneficia a las variables mientras la diferencia sea menor con la *coincidencia cercana* y mayor con la *discrepancia cercana*. Puntajes negativos se interpretan como una mala capacidad de la variable independiente para explicar la predicción correspondiente por lo que son descartados. Para corroborar los resultados obtenidos por este algoritmo, se graficaron histogramas correspondientes a las categorías extremas de la clasificación, es decir bajo y alto potencial de agua subterránea, para todas las variables. Una

variable se considera más relevante para el desempeño del modelo mientras menos superposición exista entre los histogramas. Mientras estos se encuentren más separados, se puede afirmar que una determinada variable es más útil para delimitar una categoría de otra. Posteriormente, para el análisis de importancia de variables dentro de los modelos construidos, se emplearon otras implementaciones diferentes a Relief-F especializadas de acuerdo con la tipología del algoritmo.

4.2. Construcción y evaluación de los modelos de clasificación

Para la construcción de los modelos de clasificación de aprendizaje de máquina se empleó el lenguaje de programación Python 3.6.7 en el entorno *Google Colab*. Los modelos de bosques aleatorios y redes neuronales fueron construidos con los paquetes scikit-learn 0.23.0 y Keras (ejecutado con Tensorflow 2.0 como *backend*) respectivamente.

4.2.1. Segmentación geográfica

Dependiendo de los datos seleccionados para entrenamiento y validación por su ubicación, se consideraron dos tipos de modelos en la presente investigación.

Modelos sin segmentación geográfica

Considerando la diversidad geográfica existente en el Perú y que la mayoría de los pozos con caudal conocido se encuentran en la región costa, inicialmente se optó por construir y validar los modelos únicamente en base a la información de esta región. Luego, este modelo fue utilizado para caracterizar los pozos de la sierra y de la selva con el objetivo de verificar la aplicabilidad de este en otras regiones. Estudios previos indican que los modelos de clasificación de agua subterránea construidos con algoritmos de aprendizaje de máquina de bosques aleatorios pueden caracterizar grandes extensiones geográficas adecuadamente a pesar de que el inventario de pozos sea escaso o inexistente (Moghaddam et al., 2020).

Modelos con segmentación geográfica

Una estrategia adicional abordada para incrementar el desempeño de los modelos en el contexto diverso del territorio peruano consistió en crear modelos especializados en regiones más reducidas. En otras palabras, estos fueron entrenados, validados y evaluados únicamente con la información de los pozos más cercanos entre sí. Este procedimiento se justifica en la primera ley de la geografía de Tobler, la cual establece que emplazamientos cercanos tienden a

presentar características similares. En este sentido, se optó por utilizar el método de agrupamiento por K-medias para dividir los pozos disponibles de la costa en tres grupos (costa sur, costa centro y costa norte) donde se minimice la diferencia entre las coordenadas dentro de cada grupo para garantizar cercanía geográfica. Este algoritmo de aprendizaje automatizado no supervisado se inicializa creando tres puntos al azar respecto a los cuales se calcula la distancia euclidiana para los valores de latitud (y_i) y longitud (x_i) de todos los datos. Luego, se clasifica cada dato de acuerdo con el punto donde se observó la mínima distancia. Seguidamente, los puntos inicialmente creados al azar se reemplazan por los centroides de los datos de cada uno de los tres conjuntos recientemente etiquetados. Concluido este proceso, se itera el procedimiento para optimizar la clasificación. La formulación matemática de la distancia euclidiana para este caso se presenta en la ecuación 4.10.

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{4.10}$$

Cabe resaltar que segmentos correspondientes a la región sierra y selva no fueron contemplados en la investigación debido a la escasa cantidad de pozos con información documentada. La disposición espacial de los tres grupos generados por K-medias se presenta en la figura 4.12. La finalidad del proceso de clusterización consistió en agrupar datos geográficamente semejantes para la creación de modelos especializados regionalmente en visas de obtener un mejor desempeño en comparación a un modelo único para toda la costa.

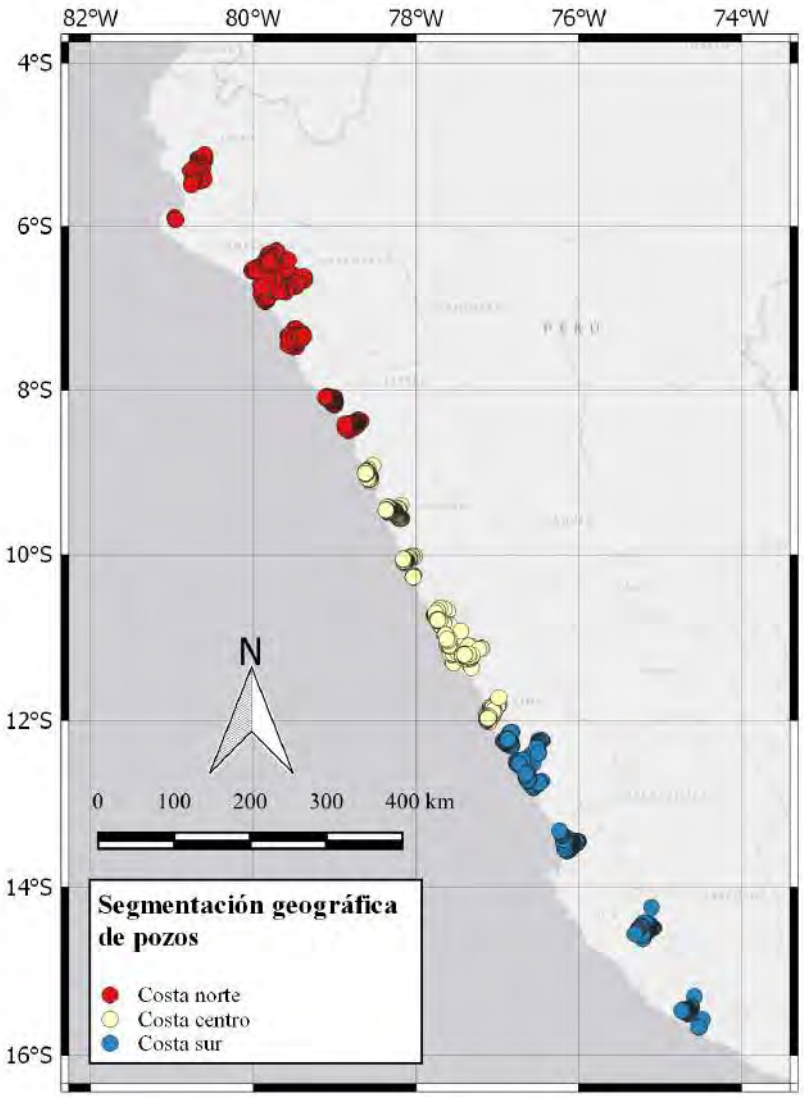


Figura 4.12. Clasificación de pozos en tres grupos de acuerdo con su cercanía geográfica empleando el algoritmo de K-medias.

4.2.2. Definición de métricas de desempeño

Para evaluar objetivamente el desempeño de los modelos de bosques aleatorios y redes neuronales construidos en la presente investigación y poder compararlos con el estado del arte, se eligieron tres métricas de desempeño: exactitud, puntaje F1 y el área bajo la curva de características operativas de receptor (AUC). Todos estas se deducen a partir de la matriz de confusión, la cual es una herramienta en la cual se puede visualizar los resultados de un modelo de clasificación. En la presente investigación, como máximo se manejan tres posibles categorías: A (potencial alto de agua subterránea), B (potencial moderado de agua subterránea) y C (potencial bajo de agua subterránea). En base a estos datos, la matriz de

confusión para todo modelo generado con estas tres categorías corresponde al esquema presentado en la tabla 4.5. Los datos de color rojo corresponden a la cantidad de elementos correctamente clasificados. Todas las otras combinaciones corresponden a discrepancias entre las clases reales y clases predichas por el modelo.

Tabla 4.5. Matriz de confusión genérica considerando tres categorías de potencial de agua subterránea.

		Clase predicha		
		C	**B**	**A**
Clase real	**C**	CC	CB	CA
	B	BC	BB	BA
	A	AC	AB	AA

Exactitud

Métrica empleada por Spiegel (2017) en la clasificación de potencial de agua subterránea de Zambia. Esta se caracteriza por su popularidad y facilidad de interpretación. La exactitud de un modelo se define como el cociente entre los elementos correctamente clasificados y el total de elementos. La ecuación 4.11 describe cómo se calcula esta métrica en la presente investigación en base a los elementos de la matriz de confusión

$$Exactitud = \frac{AA + BB + CC}{AA + AB + AC + BA + BB + BC + CA + CB + CC} \tag{4.11}$$

Los posibles puntajes de exactitud para un modelo de clasificación oscilan entre 0 y 1, siendo 0 el valor que caracteriza un peor desempeño y 1 el mejor. Sin embargo, la exactitud no basta para juzgar del desempeño adecuado de los modelos debido a un fenómeno conocido como la "paradoja de la exactitud". Para el presente caso, un modelo que clasifica indiscriminadamente a todos los pozos como de categoría C es intuitivamente un modelo mediocre. Sin embargo, obtendría una exactitud alta si los datos de evaluación se encuentran principalmente compuestos por pozos de tipo C.

Puntaje F1

Métrica creada para resolver el paradigma de la exactitud midiendo el balance entre la precisión y la exhaustividad de un modelo afín de caracterizar mejor su desempeño. Abordando a la categoría A como ejemplo, la definición matemática de precisión y exhaustividad se presenta en las ecuaciones 4.12 y 4.13.

$$Precisión = P = \frac{AA}{AA + BA + CA} \tag{4.12}$$

$$Exhaustividad = E = \frac{AA}{AA + AB + AC} \tag{4.13}$$

Interpretando estas fórmulas, mientras la exhaustividad evalúa qué tan correcto es el modelo (tasa de elementos correctamente clasificados como A sobre el total de elementos que realmente son de clase A), la precisión cuantifica la relevancia del modelo (tasa de elementos correctamente clasificados como A sobre el total de elementos correcta e incorrectamente clasificados como A). Luego, el puntaje F1 se define como la media armónica entre ambos valores como se puede apreciar en la ecuación 4.14.

$$F1 = \frac{2PE}{P + E} \tag{4.14}$$

Puesto que el presente es un problema de clasificación multiclase con clases no perfectamente uniformes, el puntaje F1 se calculó como el promedio ponderado de los puntajes correspondientes a cada una de las clases A, B y C. Para compensar las irregularidades existentes entre el número de elementos de cada categoría, el puntaje F1 computó empleando la ecuación 4.15.

$$Puntaje\ F1\ = \frac{AF1_A + BF1_B + CF1_C}{A + B + C} \tag{4.15}$$

Donde los valores de A, B y C corresponden al número de elementos correspondientes a las categorías del mismo nombre.

AUC

Métrica seleccionada puesto que es la más común en el estado del arte para la evaluación de modelos de clasificación de potencial de agua subterránea (Lee et al., 2018, 2012; Moghaddam et al., 2020; Naghibi et al., 2016; Rahmati et al., 2016). A diferencia de los puntajes anteriormente mencionados, el AUC es un valor independiente del umbral de decisión adoptado en la clasificación. El umbral se define como la probabilidad mínima para que un elemento sea clasificado en una categoría determinada y tiene por defecto un valor de 0.5. Poniendo en contexto este concepto dentro de la presente investigación, al evaluar la clasificación de un pozo a la categoría A de alto potencial, la probabilidad de pertenencia

computada por el modelo debe ser superior a 0.5 para clasificarlo en esta categoría. Para calcular el AUC de una categoría en un problema de clasificación multiclase, es necesario computar la tasa de verdaderos positivos (también conocida como sensitividad) contra los falsos positivos (también conocida como 1-especificidad) para diferentes valores de umbral y graficándolos en los ejes de las ordenadas y abscisas respectivamente. A este gráfico se le denomina curva ROC. Un modelo perfecto para todo umbral presenta constantemente una tasa de verdaderos positivos de 1 (siempre acierta) sin importar la tasa de falsos positivos y por ende, un área bajo la curva de 1. Un ejemplo de curva ROC perfecta e imperfecta se presenta en la figura 4.13.

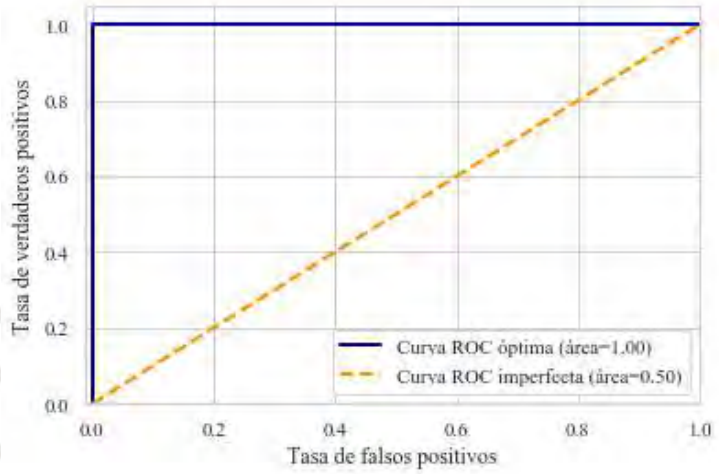


Figura 4.13. Ejemplo de curvas ROC perfecta e imperfecta para modelos de clasificación.

En el contexto de la presente investigación, para los modelos de clasificación de tres clases se utilizó una perspectiva "uno contra todos". Es decir, para cada categoría un valor de 0 de probabilidad se interpretó como pertenencia a cualquiera de las dos clases restantes. Asimismo, para analizar el desempeño global del modelo, se calculó una curva ROC promedio utilizando los resultados de las tres categorías. Esta también se denomina curva ROC micro promediada puesto que toma en cuenta la existencia de clases no balanceadas como es el presente caso de estudio. Por otro lado, para los modelos de clasificación binaria (solo dos categorías), bastó con construir una única curva ROC dado que la clasificación negativa de un elemento a la categoría A (alto potencial) automáticamente se interpreta como una clasificación de dicho elemento a la categoría B (bajo potencial). Tanto en el caso de clasificación binaria como multiclase, se determinó cuál es el valor del umbral óptimo que permite maximizar la tasa de verdaderos positivos y minimizar la tasa de falsos positivo utilizando el índice de Youden. Este se define en la ecuación 4.16.

$$J_i = E_i + S_i - 1 \tag{4.16}$$

Donde J_i, E_i y S_i corresponden al índice de Youden, la especificidad y la sensitividad para el umbral iésimo. Por definición, el umbral que obtenga el mayor índice de Youden se define como el umbral óptimo de una categoría (Ruopp, Perkins, Whitcomb, y Schisterman, 2008).

4.2.3. Construcción y evaluación de los modelos de bosques aleatorios

Todos los modelos de bosques aleatorios generados en la presente investigación se construyeron empleando un 80 % de los datos disponibles para la validación cruzada, y 20 % para la evaluación. En el proceso de validación cruzada se optó por dividir los datos en 4 grupos de modo que la cantidad de datos de evaluación y validación sea la misma. Debido a la diferencia en el orden de magnitud de las variables independientes, se empleó un escalamiento estándar para que todas presenten una distribución uniforme donde la media sea de 0 y la desviación estándar de 1. Si bien los modelos de bosques aleatorios no son afectados por la diferencia de escala de las variables independientes, se optó por este procedimiento con el objetivo de que la base de datos empleada en esta sección coincida con la de redes neuronales descrita posteriormente.

Modelamiento

Los hiperparámetros de cada modelo fueron seleccionados con el objetivo de maximizar el desempeño promedio obtenido durante el proceso de validación cruzada. En este sentido, se optimizó el número de árboles, el criterio de separación de nodos, la profundidad de árboles, el número mínimo de muestras para separar nodos, el número mínimos de muestras en hojas y el número máximo de variables consideradas al separar nodos.

Análisis de importancia de variables

Para determinar qué variables son de mayor relevancia en cada uno de los modelos bosques aleatorios, se utilizó la implementación “Feature Importance” nativa de scikit-learn. Esta se relaciona con la variación del grado de impureza asociada a cada nodo (ya sea por índice de Gini o entropía de la información) ponderado por la probabilidad de alcanzar ese nodo. De esta manera, el análisis de importancia toma en cuentan dos factores: la capacidad de los nodos para dividir a los datos de manera más diferenciada (Breiman, Friedman, Olshen, y Stone, 2017) y la frecuencia de uso de las variables en los árboles de decisión. Para efectuar este análisis, fue necesario computar la importancia de cada nodo en el modelo, asociar estos resultados a la variable correspondiente y finalmente hallar el promedio de todos los árboles de decisión. En términos matemáticos, la importancia de un nodo que se deriva en dos nodos hijos se define en

la ecuación 4.17.

$$I_i = w_i C_i - w_i^1 C_i^1 - w_i^2 C_i^2 \tag{4.17}$$

Donde I_i es la importancia asociada al nodo, w_i es la proporción de elementos que llegan al nodo correspondiente a la variable iésima y C_i es el grado de impureza del nodo. Asimismo, w_i^1 y w_i^2 son las proporciones de elementos que llegan a los nodos hijos, y C_i^1 y C_i^2 son los grados de impureza asociados a estos últimos. Luego, se determinó el grado de importancia de una variable determinada dividiendo la suma de la importancia de todos los nodos asociados a ella entre la suma de la importancia del total de nodos en el árbol de decisión. Estos datos se normalizaron para que la suma de importancias de variables dentro de un árbol sea equivalente a 1. Finalmente, la importancia de cada variable a nivel de bosque aleatorio corresponde al promedio aritmético normalizado de las importancias de las variables en cada uno de los árboles.

4.2.4. Construcción y evaluación de los modelos de redes neuronales

Similarmente al caso de los bosques aleatorios, para los modelos de redes neuronales también se subdividió los datos disponibles en dos grupos: 80 % para validación cruzada y 20 % para evaluación. De igual manera, se aplicó escalamiento estándar para todas las variables independientes. Sin embargo, a diferencia del caso anterior, todas las variables categóricas fueron transformadas utilizando el procedimiento de *One-hot Encoding*. De este modo, se creó una variable auxiliar para cada categoría donde los valores posibles son de 0 o 1. Por ejemplo, para la variable "Tipo de suelo", se crearon cinco variables auxiliares correspondientes a los cinco tipos de suelos presentes en la base de datos (suelo franco arcilloso arenoso, franco arcilloso, franco arenoso, arcillo arenoso, arcilloso y franco). Luego, para un determinado pozo, un valor de 1 correspondería al tipo de suelo existente mientras que valores de 0 serían asignados a las cuatro variables auxiliares restantes. Este procedimiento es imperativo en redes neuronales puesto que este algoritmo solo puede procesar datos numéricos. Incluyendo variables numéricas y categóricas auxiliares, se tuvieron un total de 49 variables en la nueva base de datos para redes neuronales.

Para el aprendizaje del modelo por retropropagación, se empleó como optimizador al denominado Estimación Adaptativa de Momento (*Adam*) y como función de pérdida a la entropía cruzada categórica para cada una de las tres clases definidas ya que se trata de un problema de clasificación multiclase. La entropía cruzada para un problema multiclase se define en la ecuación 4.18.

$$L(y_{ij}, \hat{y}_{ij}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} log(\hat{y}_{ij}) \tag{4.18}$$

Donde y_{ij} adopta un valor de 1 si el elemento j-ésimo real pertenece a la categoría iésima o 0 en su defecto, $\hat{y}_{ij}$ es la probabilidad estimada por el modelo de que el elemento j-ésimo pertenezca a la categoría iésima, n es el número de categorías y m ese el número de elementos.

Modelamiento

La optimización de los modelos de redes neuronales se efectuó modificando el número de capas ocultas, el número de neuronas en las capas ocultas, las épocas de entrenamiento y la inclusión de regularizadores L1, L2 y/o capas de descarte. Con el objetivo de analizar el comportamiento del modelo a lo largo de las épocas de aprendizaje, se construyeron gráficos de pérdida y de exactitud para los modelos. En todas las iteraciones se consideraron funciones de activación de tipo unidad de rectificación linear (*Relu*) para las capas de entrada y oculta. En lo que respecta a la capa de salida, se optó por emplear la función de activación *Softmax* ya que este permite obtener las probabilidades de correspondencia a las tres categorías evaluadas normalizadas a 1 para cada uno de los elementos de entrenamiento, validación y evaluación. Esto facilitó tanto la interpretabilidad de los resultados como la construcción de las curvas ROC.

Análisis de importancia de variables

Para determinar cuáles son las variables más relevantes en el desempeño del modelo de redes neuronales se optó por el Análisis de Importancia de Redes Neuronales Artificiales basado en Varianza (VIANN por sus siglas en inglés). Este método otorga un mayor nivel de importancia a las variables que presentan mayores pesos al final del entrenamiento y cuyos pesos en la neurona de la capa de entrada correspondiente sufren mayores cambios durante la etapa de entrenamiento. Se eligió este tipo de evaluación de importancia puesto que su correcta funcionalidad ha sido verificada con problemas de clasificación binaria, multiclase y regresión con elementos de bases de datos públicas de *scikit learn*. Asimismo, se ha reportado que los resultados de este análisis son compatibles con el análisis de importancia efectuado en modelos de bosques aleatorios (Rebelo de Sá, 2019).

La formulación matemática del puntaje VIANN se presenta en la ecuación 4.19.

$$VIANN(A_s) = \sum_{i=1}^{n} Var(w_{s,i})|w_{s,i}| \tag{4.19}$$

Donde A_s es la variable independiente cuya importancia se está examinando, i es el número de nodos en la primera capa oculta y $w_{s,i}$ es el peso existente entre la neurona iésima y la neurona de la variable independiente en la capa de entrada.

Capítulo V

Resultados y discusión

5.1. Preprocesamiento de datos

5.1.1. Análisis estadístico de variables

Variables continuas

Los resultados obtenidos para las variables topográficas se presentan en la tabla 5.1. Estos exceptúan a la variable categórica de relieve. Se puede observar que la variable elevación es sumamente dispersa. En contraste, la variable aspecto es la más uniforme de esta categoría.

Tabla 5.1. Análisis estadístico de las variables topográficas continuas.

	Elevación (m)	**Pendiente (°)**	**Aspecto (°)**
Media	141.810	2.725	183.725
Desviación estándar	340.614	2.636	99.268
Coeficiente de var.	240.19 %	96.72 %	54.03 %
Mínimo	0.000	0.927	0.000
Cuartil 1	15.000	1.327	108.315
Cuartil 2	47.000	2.084	191.378
Cuartil 3	114.000	2.995	270.000
Máximo	3940.000	41.038	350.332

Los resultados del análisis de las variables hidrológicas se presentan en la tabla 5.2. Se observa una alta dispersión en las variables de distancia a ríos y densidad de drenaje. En el caso de las variables relacionadas a precipitación, también se observa poca uniformidad en los datos, siendo el coeficiente de variación siempre superior a 100 %.

Tabla 5.2. Análisis estadístico de las variables hidrológicas continuas no relacionadas a la precipitación.

	D. de drenaje (m^{-1})	**D. a ríos (m)**	**Ev. (kg/m^2)**	**NDWI**	**SPI**	**TWI**
Media	55.677	12951.540	196.376	0.112	4.698	11.258
Desviación estándar	48.024	111051.100	59.925	0.059	2.817	2.603
Coeficiente de var.	86.25 %	857.44 %	30.52 %	53.04 %	59.96 %	23.12 %
Mínimo	0.000	0.000	50.073	-0.118	2.016	7.132
Cuartil 1	0.000	780.500	155.690	0.077	2.825	9.458
Cuartil 2	64.037	2231.000	199.194	0.110	3.521	10.263
Cuartil 3	81.377	7933.000	234.350	0.144	5.384	11.859
Máximo	639.646	1384048.000	320.355	0.429	13.747	19.708

Tabla 5.2 (cont.). Análisis estadístico de las variables hidrológicas continuas relacionadas a la precipitación en metros.

	Histórica	**Mes de medición**	**Año de medición**	**Intervalo de 5 años**
Media	0.0208	0.0140	0.0170	0.0140
Desviación estándar	0.0233	0.0225	0.0189	0.0201
Coeficiente de var.	111.66 %	160.86 %	111.58 %	143.73 %
Mínimo	0.0043	0.0001	0.0033	0.0006
Cuartil 1	0.0114	0.0029	0.0076	0.0028
Cuartil 2	0.0177	0.0085	0.0133	0.0101
Cuartil 3	0.0200	0.0159	0.0193	0.0176
Máximo	0.2437	0.2287	0.2153	0.2080

En el caso de las variables pedológicas y geológicas, se omitió a las variables categóricas de formación geológica y tipo de suelo para el análisis. Tal y como se presenta en la tabla 5.3, se puede observar que estas variables son de manera general menos dispersas que las anteriores mostradas con excepción de la distancia a fallas. Esto indicaría que el contenido de arcilla, arena y agua es relativamente uniforme entre las localidades donde hay pozos con caudal conocido.

Tabla 5.3. Análisis estadístico de las variables geológicas continuas

	D. a fallas (m)	**C. de arena (%)**	**C. de arcilla (%)**	**C. de agua (%)**
Media	562541.700	50.248	27.707	24.151
Desviación estándar	670357.600	6.289	5.704	7.978
Coeficiente de var.	119.17 %	12.52 %	20.59 %	33.03 %
Mínimo	0.000	31.000	10.000	1.000
Cuartil 1	8847.500	46.000	23.000	19.000
Cuartil 2	15702.000	49.000	29.000	27.000
Cuartil 3	1376328.000	56.000	32.000	29.000
Máximo	1382875.000	78.000	41.000	50.000

Respecto a la variable de NDVI y a la variable dependiente (caudal), el análisis se expone en la tabla 5.4. A partir de estos valores, cabe mencionar que la variable de NVDI posee una dispersión baja en comparación al resto de las variables examinadas. En lo concerniente al caudal, se aprecia una dispersión considerable. Esto se debe a que el 75 % de los valores son inferiores a 18 lps. Sin embargo, el valor máximo de 100 lps excede de sobre manera (casi 6 desviaciones estándar) al tercer cuartil. Para visualizar de mejor manera la considerable dispersión de los datos, se graficó el histograma correspondiente al caudal de los pozos en la figura 5.1. Se puede observar que la mayoría se concentra en rangos de bajo caudal. A partir de los 10 lps, el número de pozos cae considerablemente. Asimismo, debido a la presencia de caudales repetidos, la categorización por percentiles no se realizó de manera perfectamente uniforme. Como se aprecia en la figura 5.2, existe un menor número de pozos correspondiente a la categoría B, sesgando el desempeño en contra de esta categoría.

Tabla 5.4. Análisis estadístico de la variable NDVI y el caudal de los pozos

	NDVI	**Caudal (lps)**
Media	0.283	13.586
Desviación estándar	0.119	14.269
Coeficiente de var.	41.95 %	105.03 %
Mínimo	0.034	0.500
Cuartil 1	0.170	5.000
Cuartil 2	0.279	10.000
Cuartil 3	0.378	18.000
Máximo	0.616	100.000

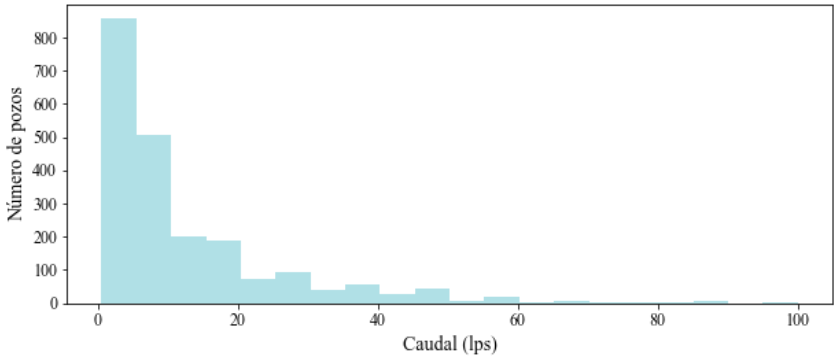


Figura 5.1. Histograma de caudal de los pozos recopilados.

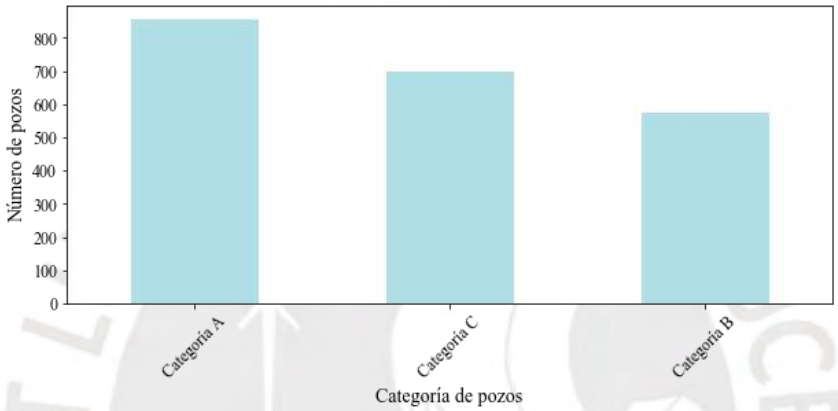


Figura 5.2. Distribución de pozos de acuerdo con las categorías definidas de potencial de agua subterránea.

Para las variables categóricas de relieve, tipo de suelo y formación geológica, los resultados del análisis de frecuencias se presentan en las tablas 5.5, 5.6 y 5.7 respectivamente.

Tabla 5.5. Análisis de frecuencias de las clases asociadas a la variable relieve.

Categoría	**Frecuencia absoluta**	**Frecuencia relativa**
Pendiente inferior (llana)	1261	59.29 %
Pendiente inferior (cálida)	399	18.76 %
Valle	226	10.63 %
Pendiente superior (llana)	200	9.40 %
Pendiente superior (cálida)	41	1.93 %

Tabla 5.6. Análisis de frecuencias de las clases asociadas a la variable tipo de suelo.

Categoría	**Frecuencia absoluta**	**Frecuencia relativa**
SaClLo	1429	67.18 %
ClLo	393	18.48 %
SaLo	293	13.78 %
SaCl	7	0.33 %
Cl	3	0.14 %
Lo	2	0.09 %

Tabla 5.7. Análisis de frecuencias de las clases asociadas a la variable formación geológica.

Categoría	**Frecuencia absoluta**	**Frecuencia relativa**
Qh-c	1455	68.41 %
Ki-c	275	12.93 %
Qp-c	91	4.28 %
Kis-vs	80	3.76 %
Ks-mgr,gd	59	2.77 %
JsKi-mc	48	2.26 %
Ks-and,ri	29	1.36 %
Np-vs	19	0.89 %
N-gb,di	19	0.89 %
JsKi-vs	17	0.80 %
Ks-di,tn,gd	15	0.71 %
Ks-tn,gd	7	0.33 %
KP-mgr,gd	4	0.19 %
Ki-m	2	0.09 %
KP-tn,gd	2	0.09 %
Ks-gb,di	1	0.05 %
PN-vs	1	0.05 %
P-c	1	0.05 %
Jm-vs	1	0.05 %
PEl-c	1	0.05 %

A partir de los resultados obtenidos, es importante resaltar que en ninguno de los casos se disponen de datos correspondientes a la totalidad de categorías. Por ejemplo, el caso de la variable relieve, solo se disponen pozos correspondientes a 5 categorías de las 15 que se pueden observar en la leyenda de la figura 4.8d. Categorías tales como "risco", "montaña", entre otras, no corresponden a ninguno de los pozos empleados para el entrenamiento. En este sentido, todo modelo que sea calibrado con estos datos no podrá ser capaz de utilizar adecuadamente las categorías no contempladas para la estimación del potencial de agua subterránea. Similarmente, para el caso de la variable tipo de suelo, solo se encuentran representadas 6 categorías (suelo franco arcilloso limoso, franco arcilloso, franco arenoso, arcillo arenoso, arcilloso y franco) de las 12 totales. Cabe resaltar que ninguno de los tipos de suelo limoso está siendo contemplado en el modelo. Finalmente, en el caso de la variable de formaciones geológicas, los pozos solo corresponden a 20 de las 105 formaciones existentes en el territorio nacional. Asimismo, otro aspecto a destacar es que en todos los casos existe una categoría con una frecuencia abrumadoramente superior en comparación a las restantes: "pendiente inferior (llana)" para la variable relieve, "suelo franco arcilloso arenoso" para la variable tipo de suelo y "Qh-c" para la variable formación geológica. Por este motivo, los modelos desarrollados en la presente investigación se encuentran parcializados a estas categorías en sus estimaciones.

5.1.2. Análisis de independencia de las variables numéricas

La matriz de correlación correspondiente a las variables numéricas se presenta en la figura 5.3. En esta, se presentan los coeficientes de correlación de Pearson correspondientes a todas las combinaciones de pares de variables presentes en la investigación, siendo los valores más cercanos a 1 o -1 indicadores de una dependencia lineal más estrecha. De acuerdo con los resultados obtenidos, las variables con mayor correlación son las de contenido porcentual de arena ("sand" en la figura), arcilla ("clay" en la figura), agua ("soilwater" en la figura) y NDVI. En todos estos casos se observa una correlación negativa superior a 0.5. Por otro lado, una correlación positiva fuerte existe entre las cuatro variedades de la variable precipitación ("lluviaprom", "precip_exacta", "precip_prom_anual" y "precip_periodo" en la figura). Similarmente, una correlación positiva alta de 0.89 existe entre las variables hidrológicas de TWI y SPI.

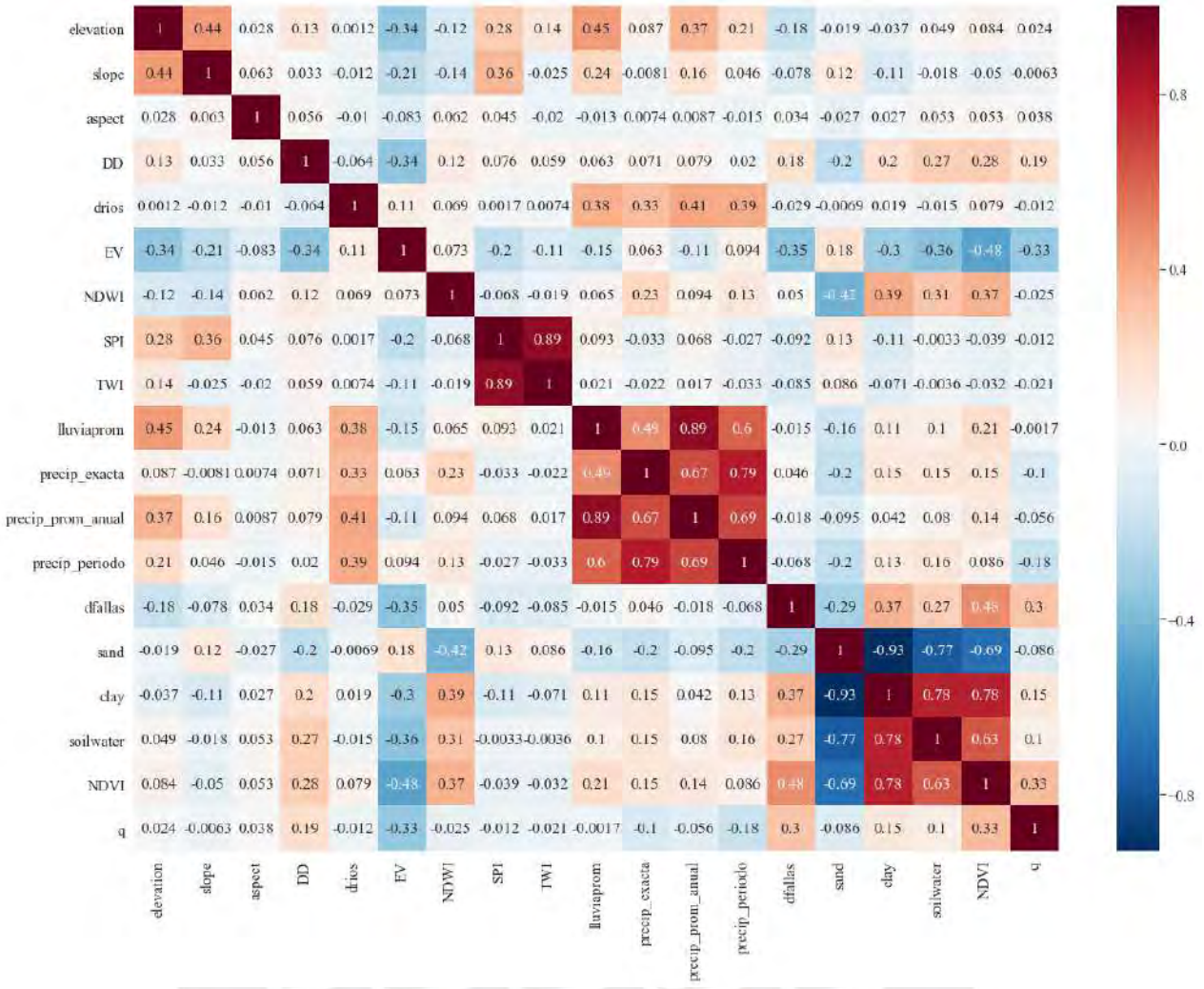


Figura 5.3. Matriz de correlación para las variables continuas de la base de datos.

La información presentada en la matriz de correlación fue complementada por el parámetro VIF. Los valores VIF obtenidos para la presente base de datos se presentan en la tabla 5.8. A partir de estos resultados, se verificó que las variables SPI, TWI, contenido porcentual de arena, contenido porcentual de arcilla, contenido porcentual de agua y NDVI podrían potencialmente ocasionar problemas de multicolinealidad en los modelos predictivos. Asimismo, cabe resaltar que se adicionó a la variable de evapotranspiración al grupo de alta correlación a causa del valor VIF obtenido de 21.56. Estas observaciones fueron tomadas en cuenta al momento de construir los modelos de clasificación con reducción de variables.

Tabla 5.8. Valores del parámetro VIF para cada una de las variables independientes continuas.

Variable	Puntaje VIF
TWI	289.17
Contenido % de arcilla	102.84
Contenido % de arena	74.72
SPI	64.12
Contenido % de agua	29.50
NDVI	25.24
Evapotranspiración	21.56
Precipitación (año)	14.23
Precipitación histórica	11.69
Pendiente	8.08
NDWI	6.51
Precipitación (+/- 5 años)	5.44
Precipitación (mes)	5.09
Aspecto	4.57
Densidad de drenaje	2.80
Distancia a fallas	2.69
Elevación	2.21
Distancia a ríos	1.35

5.1.3. Análisis preliminar de importancia de variables

En la tabla 5.9, se presentan los puntajes computados por el algoritmo Relief-F ordenados de mayor a menor para todas las variables independientes continuas. Como se puede observar, todos los puntajes obtenidos son positivos por lo que se afirma que la selección de las variables es adecuada. Asimismo, cabe resaltar que las variables de evapotranspiración, NDVI y distancia a fallas son las más relevantes para la estimación de agua subterránea de acuerdo con los resultados obtenidos. Estos resultados concuerdan con los valores obtenidos en la matriz de correlación, puesto que estas variables son las que obtuvieron un mayor coeficiente en correspondencia a la variable dependiente (“q” en la figura 5.3) con valores de -0.33, 0.33 y 0.3 respectivamente.

Tabla 5.9. Puntajes del algoritmo Relief-F para las variables independientes continuas.

Variable	**Puntaje Relief-F**
Evapotranspiración	0.1247
NDVI	0.0628
Distancia a fallas	0.0479
Contenido % de arcilla	0.0456
Contenido % de agua	0.0437
Aspecto	0.0402
SPI	0.0338
Contenido % de arena	0.0281
Densidad de drenaje	0.0269
TWI	0.0258
Elevación	0.0210
Precipitación (+/- 5 años)	0.0166
NDWI	0.0158
Precipitación (mes)	0.0126
Precipitación (año)	0.0125
Precipitación histórica	0.0121
Pendiente	0.0054
Distancia a ríos	0.0017

Estas observaciones fueron verificadas empleando los histogramas de la figura 5.4. Se puede apreciar que en el caso de las variables de evapotranspiración ("EV" en la figura) y NDVI, los histogramas correspondientes a bajo y alto potencial de agua subterránea presentan una mayor separación en comparación al resto de variables. En cambio, variables como pendiente ("slope" en la figura), SPI y TWI presentan histogramas prácticamente superpuestos por lo que se espera que su contribución al desempeño de los modelos sea menor.

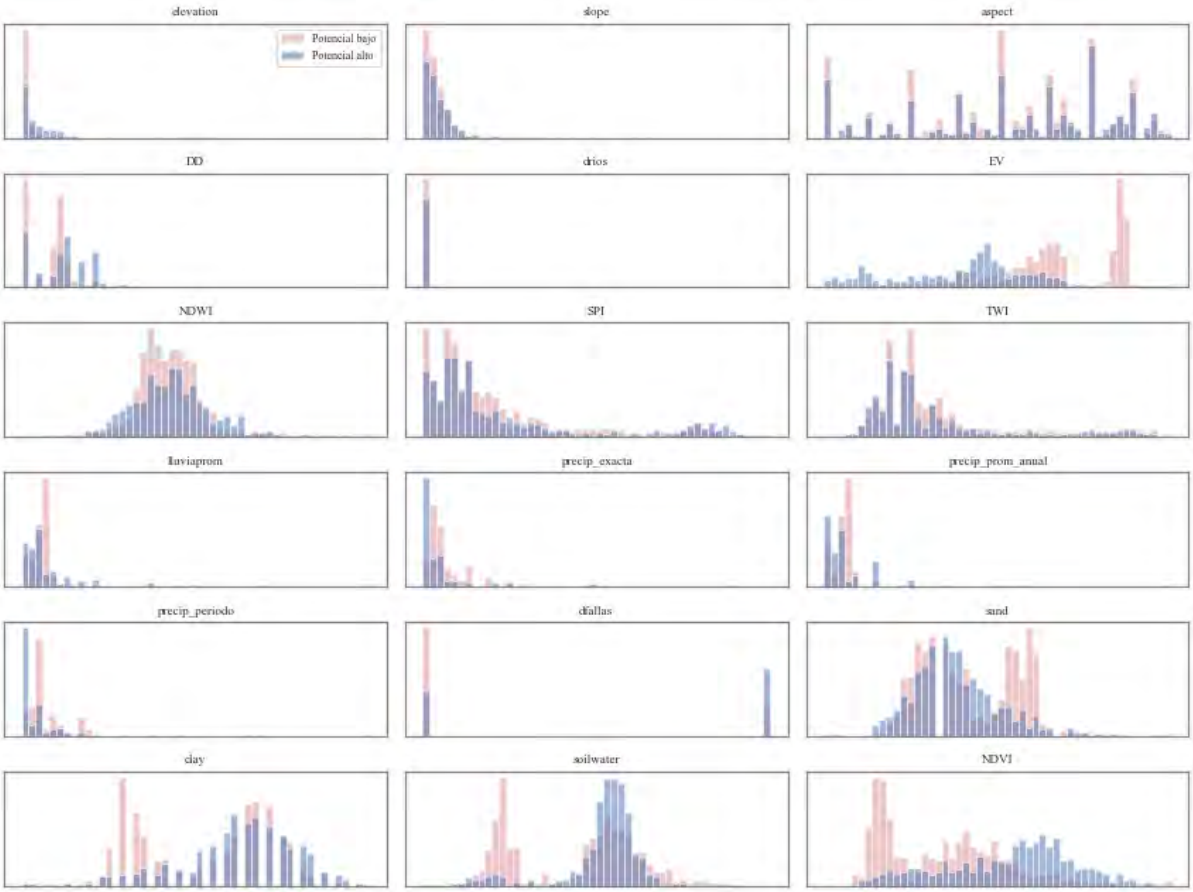


Figura 5.4. Histogramas separados de acuerdo con el potencial de agua subterránea para las variables independientes numéricas.

5.2. Construcción y evaluación de los modelos de bosques aleatorios

5.2.1. Modelos sin segmentación geográfica

Entrenamiento y validación

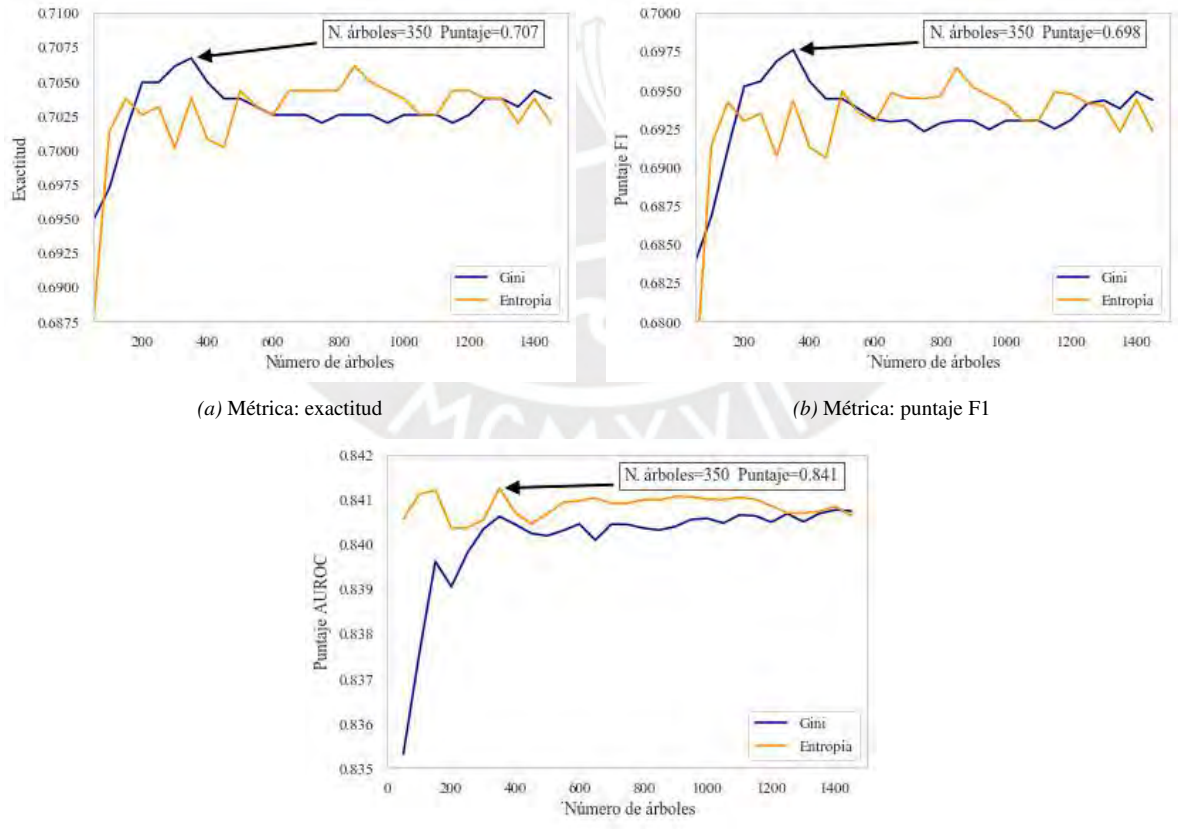
Se realizaron múltiples iteraciones del proceso de validación cruzada para encontrar los hiperparámetros que maximizaron el desempeño del modelo para las tres métricas definidas. Estos se presentan a continuación:

- Número de árboles: 350, un mayor de número produce una caída en el desempeño o un incremento marginal en comparación al incremento en el tiempo de cómputo.
- Criterio de separación de nodos: impureza de Gini, se utilizó esta función para la construcción de los árboles puesto que suponía mejores resultados en comparación a la

entropía.

- Profundidad de árboles: 8, se observó que la opción por defecto de no limitar la profundidad de los árboles impactaba negativamente al desempeño del modelo.
- Número mínimo de muestras para separar nodos: 2, valor mínimo y por defecto. Valores superiores reducen el desempeño del modelo.
- Número mínimo de muestras en hojas: 1, valor mínimo y por defecto. Valores superiores reducen el desempeño del modelo.
- Número máximo de variables al separar nodos: 7, correspondiente a la raíz cuadrada del número de variables. Alternativas típicas incluyen tomar en cuenta el logaritmo en base dos del número de variables o el total de variables. En ambos casos, se observó una reducción del desempeño.

Para analizar la variación del desempeño del modelo frente al criterio de separación de nodos y el número de árboles de decisión se preparó la figura 5.5. Se puede observar que el modelo seleccionado con impureza de Gini y 350 árboles supera al de entropía de la información en dos de tres de las métricas propuestas. Asimismo, su AUC solo es despreciablemente inferior al modelo que emplea entropía de la información.



(a) Métrica: exactitud

(b) Métrica: puntaje F1

(c) Métrica: AUC

Figura 5.5. Comparación entre los modelos de bosques aleatorios que emplean índice Gini y entropía de la información para las tres métricas definidas.

Evaluación del modelo

Al clasificar a los datos del conjunto de evaluación, el valor de exactitud obtenido fue de 0.72. Los valores de precisión, exhaustividad y el puntaje F1 correspondiente a cada clase y ponderado global para el modelo para se presentan en la tabla 5.10.

Tabla 5.10. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de bosques aleatorios sin segmentación geográfica y sin reducción de variables.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.72	0.82	0.76
B	0.66	0.45	0.53
A	0.75	0.80	0.78
Promedio ponderado	0.72	0.72	0.71

A partir de los resultados obtenidos, se puede observar que, para las clases extremas de bajo y alto potencial de agua subterránea, se obtienen los valores más favorables de precisión, exhaustividad y puntaje F1. Es decir, el modelo predice de manera adecuada estas dos categorías. No obstante, se aprecia una exhaustividad considerablemente baja para la clase B de potencial moderado de agua subterránea. Esto se debe a la menor cantidad de datos de entrenamiento disponibles de esta categoría como se observó anteriormente. A pesar de estas dificultades, el modelo obtiene un puntaje F1 ponderado de 0.71 y valores balanceados de precisión y exhaustividad. En este sentido, se puede afirmar que el modelo no recae en la "paradoja de la exactitud" mencionada anteriormente. Por último, las curvas ROC se presentan en la figura 5.6. Los resultados indican nuevamente que el modelo presenta una mejor capacidad para la clasificación de las categorías extremas, obteniendo puntajes AUC altos de 0.87 y 0.90. Agregando los valores de las curvas, se obtuvo un puntaje AUC promedio global de 0.86. Los umbrales óptimos que minimizan los falsos positivos y maximizan los verdaderos positivos son de 0.50, 0.28 y 0.40 para las categorías C, B y A respectivamente empleando el índice de Youden. En este sentido, de adoptar este umbral para la categoría A, se tendría una tasa de verdaderos positivos de 0.79 y una tasa de falsos positivos de 0.11. En otras palabras, del total de emplazamientos categorizados como A por el modelo, el 79 % sería correcto. Asimismo, el modelo sería capaz de descartar correctamente la categoría A para un emplazamiento el 89 % de veces.

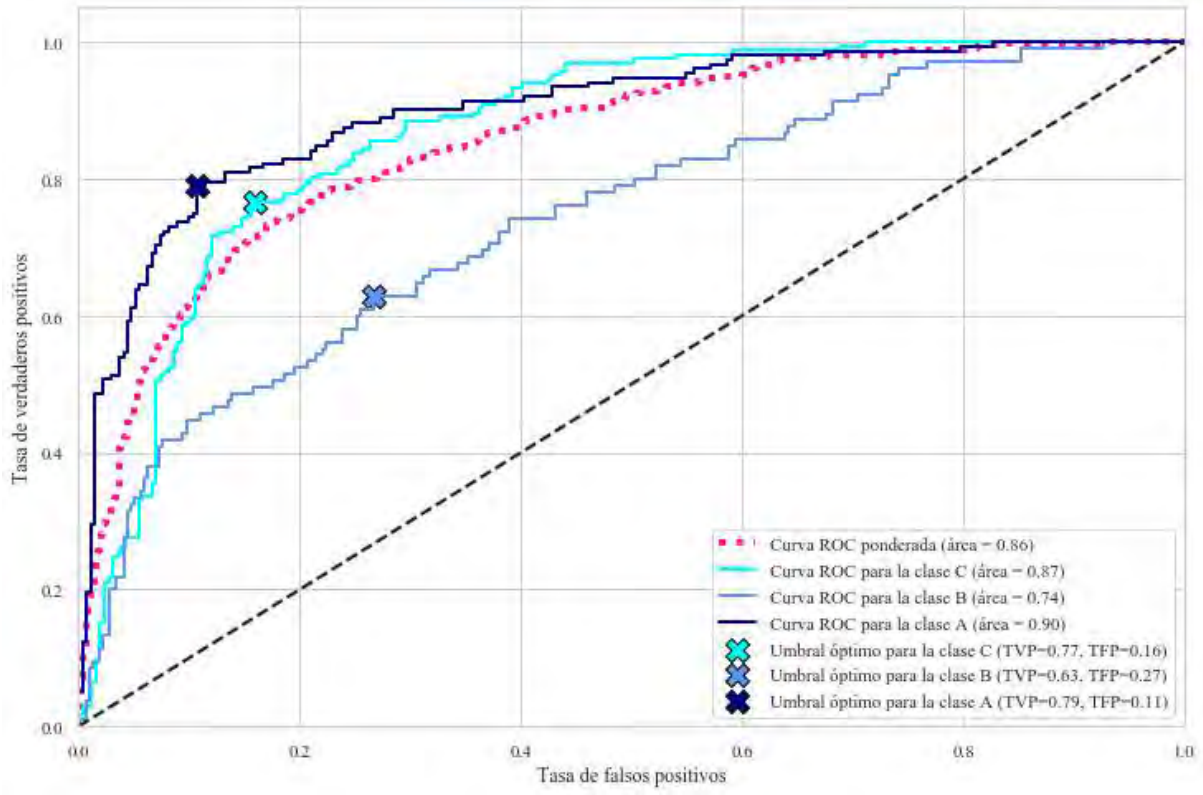


Figura 5.6. Curvas ROC correspondientes al modelo de bosques aleatorios sin segmentación geográfica con todas las variables.

Análisis de importancia de variables

En la figura 5.7, se presenta el análisis de importancia de variables ordenado de mayor a menor. Se puede observar que las variables más relevantes para el desempeño del modelo son las de evapotranspiración (EV), precipitación del mes de medición en un periodo de + y - 5 años (precip_periodo), elevación (elevation), distancia a fallas (dfallas) y NDVI. Cabe mencionar que, de este grupo, las variables de evapotranspiración, NDVI y distancia a fallas también obtuvieron los puntajes más altos del análisis preliminar de importancia con el algoritmo Relief-F (ver tabla 5.9). Sin embargo, las variables de evapotranspiración y NDVI también exhibieron posibles problemas de multicolinealidad según el puntaje VIF obtenido (ver tabla 5.8). Por otro lado, en lo referido a las variables con menor importancia en el desempeño del modelo, se encuentran principalmente las variables categóricas. En la figura 5.7, se puede observar una caída considerable de la importancia relativa a partir de la variable “Geo” correspondiente a las formaciones geológicas de los emplazamientos. Las variables subsiguientes también son categóricas y corresponden al tipo de relieve y al tipo de suelo. Por otro lado, las variables continuas con menor importancia son las de cantidad porcentual de arcilla (clay) y arena (sand), aspecto, pendiente y SPI. Estos resultados contradicen parcialmente a los puntajes Relief-F ya que las variables de contenido porcentual de arcilla, arena y aspecto se encontraban dentro de las 7 variables con mayor importancia preliminar.

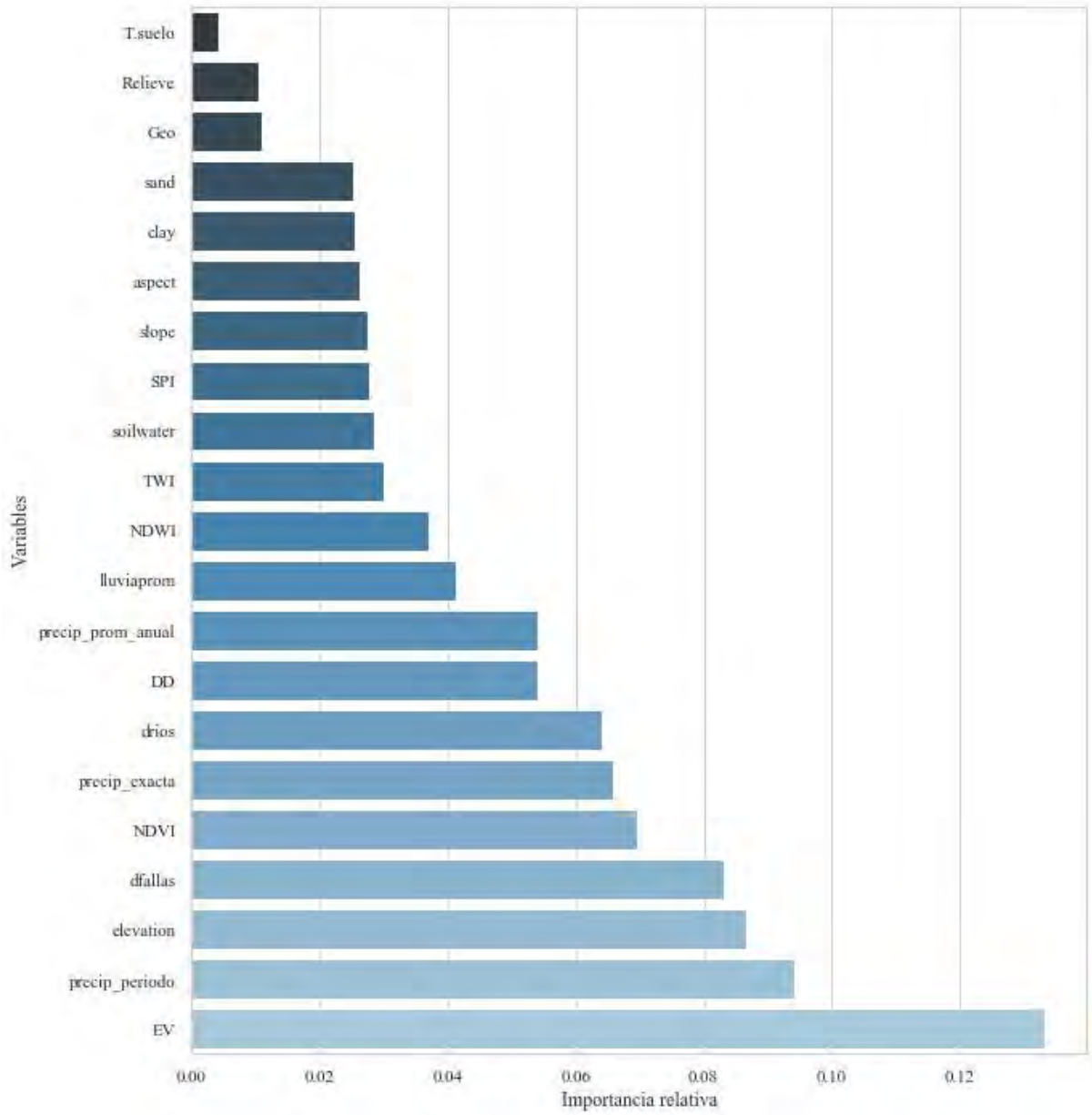
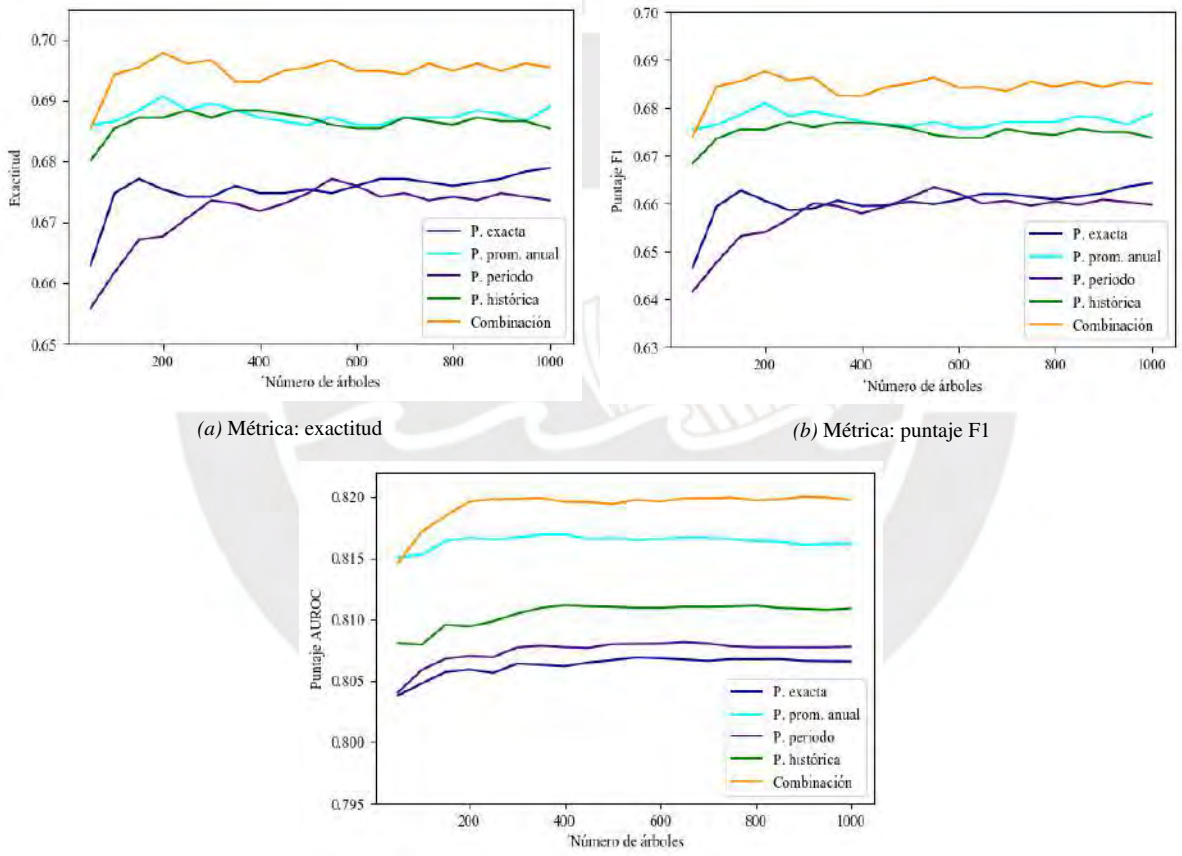


Figura 5.7. Análisis de importancia de variables correspondiente al modelo de bosques aleatorios sin segmentación geográfica y sin reducción de variables.

Reducción de variables

En base a estos resultados, se optó por construir un segundo modelo eligiendo únicamente a las variables más importantes procurando mantener el desempeño constante. De esta manera, se redujo la complejidad computacional del modelo y se simplificó el problema al reducir la necesidad de información para evaluar el potencial de agua subterránea. Además, se realizó este procedimiento para comprobar si las variables con alto puntaje VIF estaban afectando la interpretabilidad de la importancia de variables a causa del problema de multicolinealidad. En primer lugar, se eliminaron todas las variables categóricas dado que su representación no uniforme de las categorías impide contribuir al aprendizaje del modelo.

Asimismo, puesto que las variables con altos valores de VIF se pueden explicar a través de la combinación lineal de las otras variables, se eliminaron todas las que obtuvieron un puntaje mayor a 15 (TWI, contenido porcentual de arcilla y arena, SPI, contenido porcentual de agua, NDVI y evapotranspiración). Asimismo, se eliminaron las variables con los puntajes más bajos de importancia Relief-F (pendiente y distancia a ríos). De esta manera, quedaron restantes 8 variables independientes: aspecto, densidad de drenaje, elevación, NDWI, precipitación en un periodo de + y - 5 años, precipitación exacta, precipitación anual y precipitación histórica. Si bien se consideró descartar alguna de las variables de precipitación, durante la fase de validación cruzada se observó que usar simultáneamente las cuatro variables planteadas de precipitación mejoraba el desempeño del modelo para las tres métricas planteadas en comparación a solo usar una a la vez, independientemente del número de árboles. Estos resultados se observan en las gráficas de desempeño de la figura 5.8.



(a) Métrica: exactitud

(b) Métrica: puntaje F1

(c) Métrica: AUC

Figura 5.8. Análisis de importancia de los tipos de precipitación independientemente y en simultáneo para el modelo de bosques aleatorios.

Para este nuevo modelo de variables reducidas, los hiperparámetros óptimos se actualizaron a 200 para el número de árboles y 7 para la profundidad de árboles, manteniendo el resto constantes. Evaluando el desempeño del modelo con los datos de evaluación, se obtuvieron métricas cuyo valor difiere como máximo en 0.02 en comparación al modelo de 21 variables:

exactitud de 0.70, puntaje F1 ponderado de 0.69 y puntaje AUC de 0.84. El detalle de precisión y exhaustividad por categoría se presenta en la tabla 5.11. Las nuevas curvas ROC se presentan en la figura 5.9, donde los umbrales óptimos son de 0.38, 0.34 y 0.49 para las categorías C, B y A respectivamente.

Tabla 5.11. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de bosques aleatorios sin segmentación geográfica y con reducción de variables.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.70	0.82	0.75
B	0.65	0.41	0.50
A	0.73	0.78	0.76
Promedio ponderado	0.70	0.70	0.69

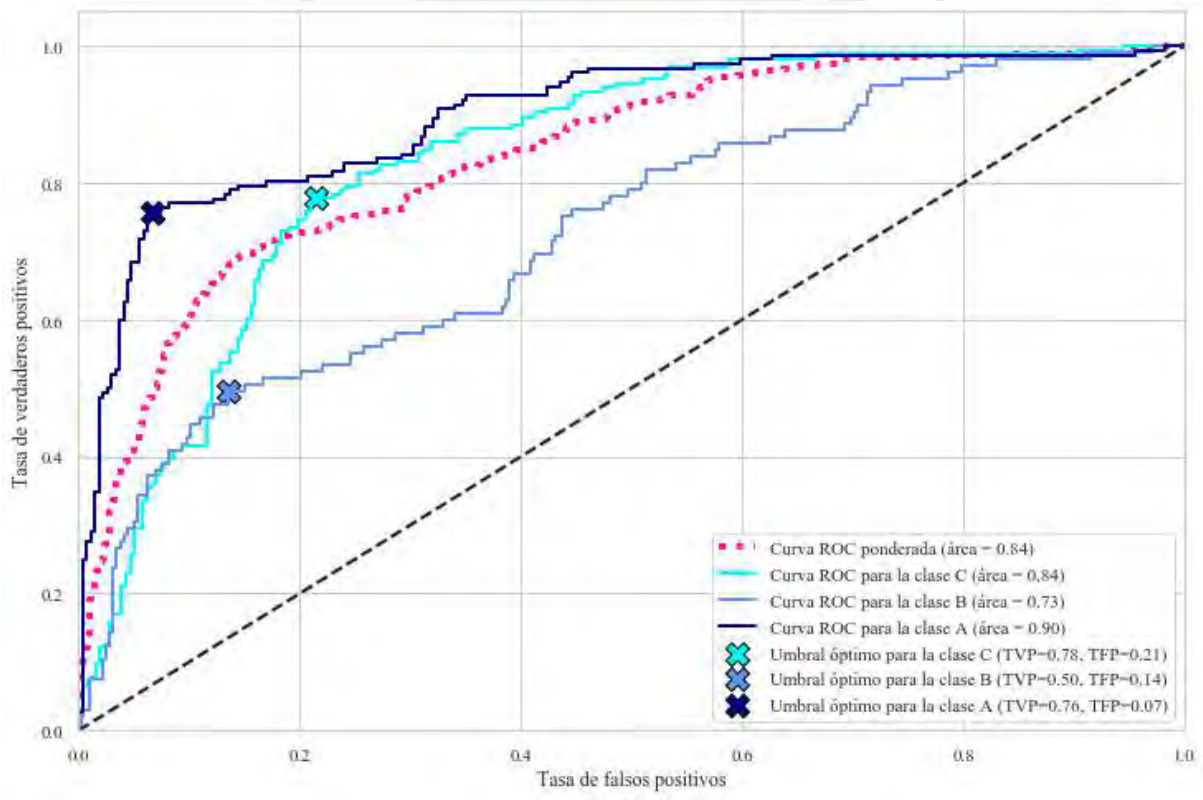


Figura 5.9. Curvas ROC correspondientes al modelo de bosques aleatorios sin segmentación geográfica y con reducción de variables.

El nuevo orden de importancia de variables se presenta en la figura 5.10. Al haber eliminado las variables con problemas de multicolinealidad en el modelo, la interpretabilidad de estos resultados no se encuentra afectada. Se puede apreciar que las variables más importantes son las de elevación y precipitación en un periodo de 5 años. En contraposición, el aspecto topográfico

y el NDWI son las variables con menor relevancia en el desempeño del modelo de bosques aleatorios.

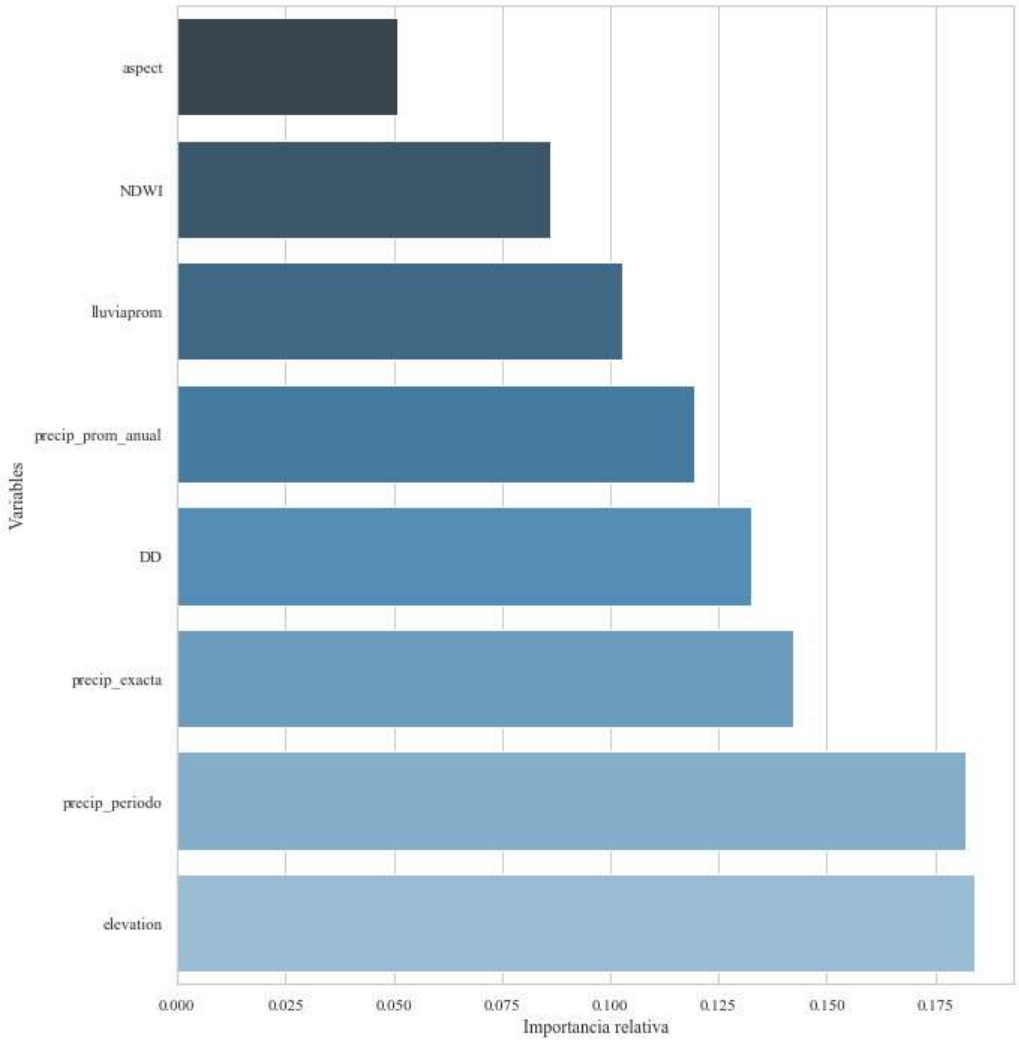


Figura 5.10. Análisis de importancia de variables correspondiente al modelo de bosques aleatorios sin segmentación geográfica y con reducción de variables.

5.2.2. Modelos con segmentación geográfica

Costa sur

Se obtuvo una exactitud de 0.77, superior a la obtenida por los modelos sin segmentación. Los valores de precisión, exhaustividad y puntaje F1 se presentan en la tabla 5.12. Similarmente al caso de la exactitud, se observa un aumento de múltiples parámetros de precisión y exhaustividad para las clases A y C. En consecuencia, el modelo para la costa sur posee un puntaje F1 ponderado de 0.73, superior al obtenido por los modelos sin segmentar. Sin embargo, se puede observar un valor mediocre en lo correspondiente a la exhaustividad de

la categoría B. Esto se debe a que además de reducirse el número de pozos disponibles para la construcción de modelos a 1292, la minoría corresponde a la categoría B, sesgando el aprendizaje del modelo. Si bien la división de categoría se volvió a realizar utilizando los nuevos percentiles 33 y 66 para intentar obtener tres categorías uniformes, la presencia de múltiples pozos con el mismo valor de caudal en la costa norte impidió generar clases balanceadas.

Tabla 5.12. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de bosques aleatorios segmentado para la costa sur.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.76	0.96	0.85
B	0.85	0.22	0.35
A	0.78	0.76	0.77
Promedio ponderado	0.78	0.77	0.73

Las curvas ROC, las AUC y los umbrales óptimos se presentan en la figura 5.11. Se aprecia un aumento del parámetro AUC promedio ponderado a 0.88. Asimismo, para las clases C, B y A, se determinaron como umbrales últimos a los valores de 0.51, 0.26 y 0.48, obteniendo las tasas de verdaderos positivos y falsos positivos presentadas en la figura. Habiendo superado a los modelos sin segmentación en todas las métricas, se puede afirmar que la estrategia de segmentación geográfica fue adecuada para la costa sur. Cabe resaltar que gracias a la disponibilidad de datos en la costa sur, fue posible mantener la división tripartita basada en percentiles para garantizar un balance de clases de caudales sin disminuir el alcance del problema.

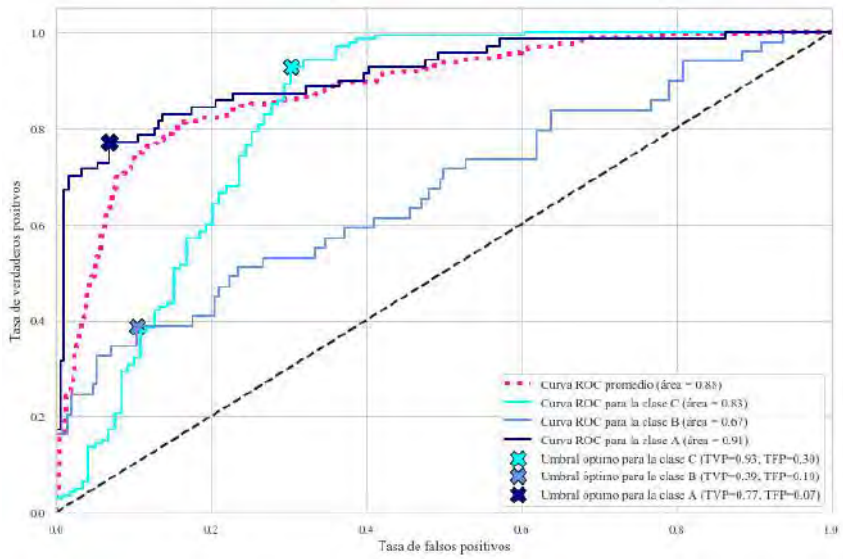


Figura 5.11. Curvas ROC correspondientes al modelo de bosques aleatorios segmentado para la costa sur.

Costa centro

De acuerdo con la segmentación propuesta para la costa centro, tan solo se dispusieron 392 pozos para la construcción del modelo. Asimismo, puesto que múltiples pozos presentaban valores de caudal idénticos o muy similares, se observó una distribución muy desigual de las categorías al considerar tres grupos de potencial de agua subterránea. Para solucionar ambos impases, se optó por emplear solo dos categorías denominadas A y B para caracterizar a los emplazamientos con potencial de agua subterránea superiores e inferiores a la mediana regional respectivamente. Tal y como se explicó en la página 27 de la subsección 4.1.1, ello significó considerar solo dos grupos balanceados que contienen a los pozos con caudales sobre (clase A) y por debajo de segundo cuartil (clase B). Bajo estas consideraciones, se obtuvo una exactitud de 0.81. Los valores de precisión, exhaustividad y puntaje F1 se presentan en la tabla 5.13.

Tabla 5.13. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de bosques aleatorios segmentado para la costa centro.

Clase	Precisión	Exhaustividad	Puntaje F1
B	0.78	0.93	0.85
A	0.88	0.64	0.74
Promedio ponderado	0.82	0.81	0.80

Como se puede apreciar en la figura 5.12, se obtuvo un área de 0.87 bajo la curva. El umbral óptimo según el índice de Youden es de 0.45, obteniendo así maximizando la tasa de verdaderos positivos a 0.73 y minimizando la tasa de falsos positivos a 0.09.

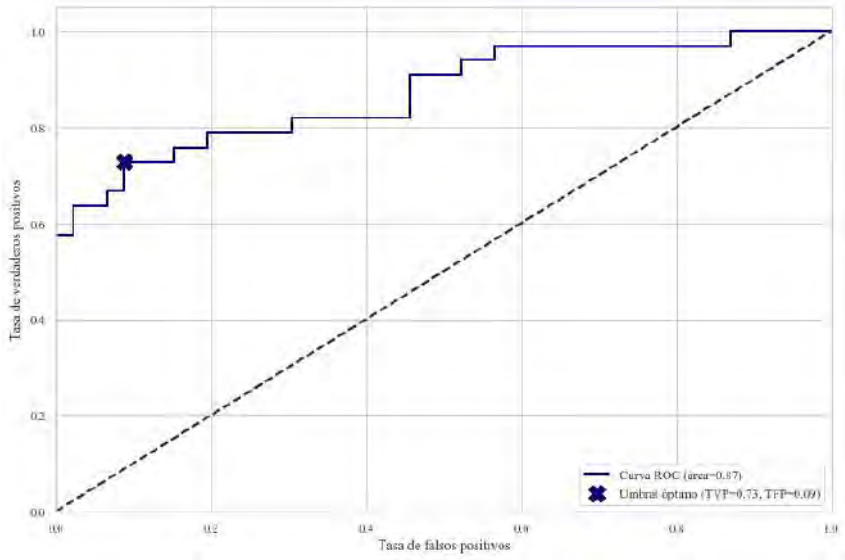


Figura 5.12. Curva ROC correspondiente al modelo de bosques aleatorios segmentado para la costa centro.

Costa norte

Un problema similar al de la costa centro se observó en la costa norte puesto que tan solo 430 pozos corresponden a esta segmentación. Por este motivo, también se optó por reducir el problema a uno de clasificación binaria para distinguir a los emplazamientos con potencial mayor (categoría A) o inferior (categoría B) a la mediana regional de 14 lps. Los valores de precisión, exhaustividad y puntaje F1 se presentan en la tabla 5.14. Se obtuvo una exactitud y un puntaje F1 ponderado de 0.78.

Tabla 5.14. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de bosques aleatorios segmentado para la costa norte.

Clase	Precisión	Exhaustividad	Puntaje F1
B	0.75	0.77	0.76
A	0.80	0.79	0.80
Promedio ponderado	0.78	0.78	0.78

La curva ROC correspondiente a la categoría A de alto potencial se presenta en la figura 5.13, se obtuvo un área de 0.84 bajo la curva. El umbral óptimo según el índice de Youden es de 0.56, obteniendo así maximizando la tasa de verdaderos positivos a 0.72 y minimizando la tasa de falsos positivos a 0.13.

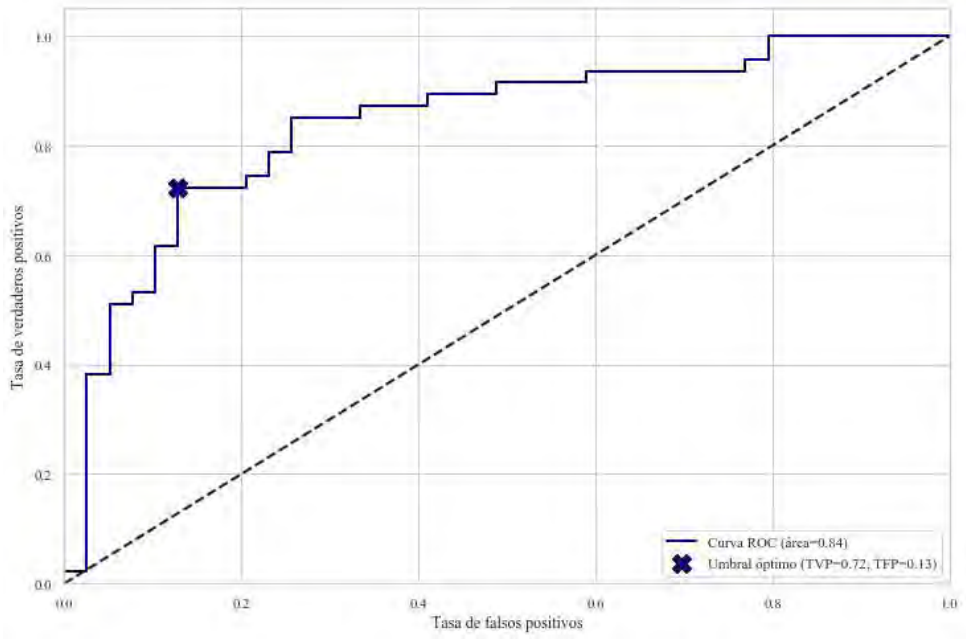


Figura 5.13. Curva ROC correspondiente al modelo de bosques aleatorios segmentado para la costa norte.

5.3. Construcción y evaluación de los modelos de redes neuronales

5.3.1. Modelos sin segmentación geográfica

Arquitectura

La arquitectura que supuso un mejor resultado para este primer modelo consistió en una red con 1 capa oculta de 100 neuronas. Se observó un sobreentrenamiento considerable al entrenar el modelo de redes neuronales sin regularizadores. Este comportamiento se puede observar en la figura 5.14, donde se presenta la variación de los valores de pérdida y de exactitud conforme transcurren las épocas de entrenamiento del modelo sin regularizadores para uno de los grupos de la validación cruzada. Se puede observar que la función de pérdida se reduce constantemente para el conjunto de entrenamiento, mientras que aumenta para el conjunto de validación. Por otro lado, un comportamiento inverso se observa para la exactitud. En este sentido, se puede afirmar que el modelo de redes neuronales sin regularizadores se adapta en exceso a los datos de entrenamiento, perjudicando su capacidad de clasificación al exponerlo a nuevos datos.

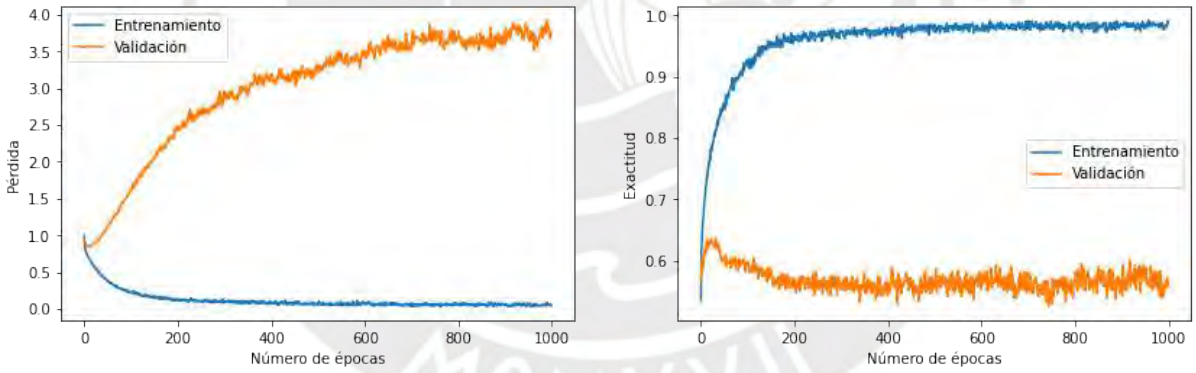
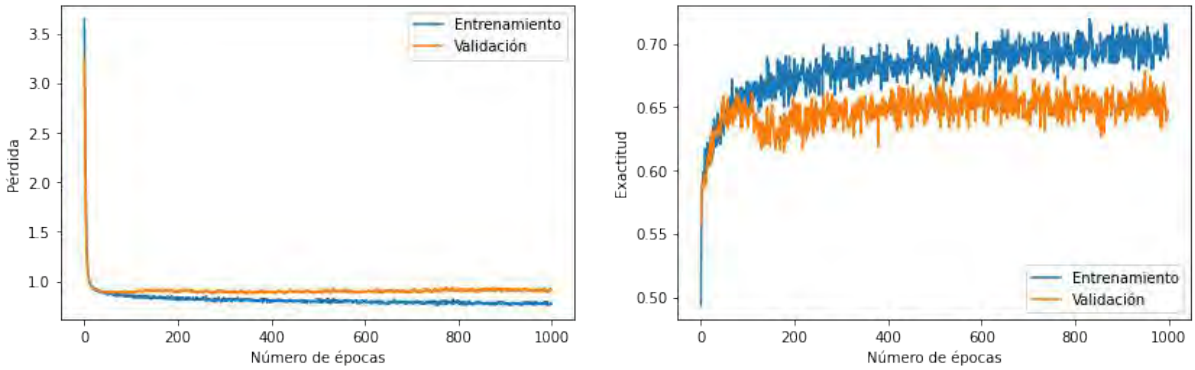
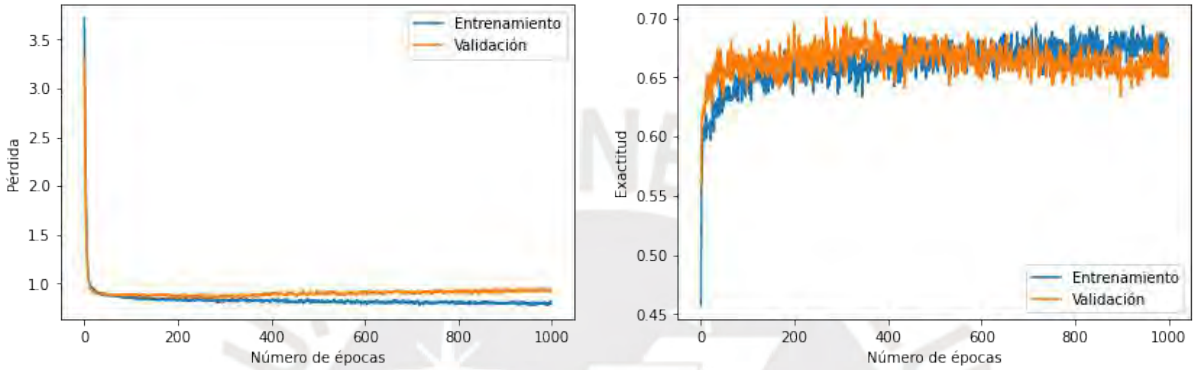


Figura 5.14. Variación de la pérdida y exactitud del modelo de redes neuronales sin segmentación geográfica y sin regularización durante la validación cruzada.

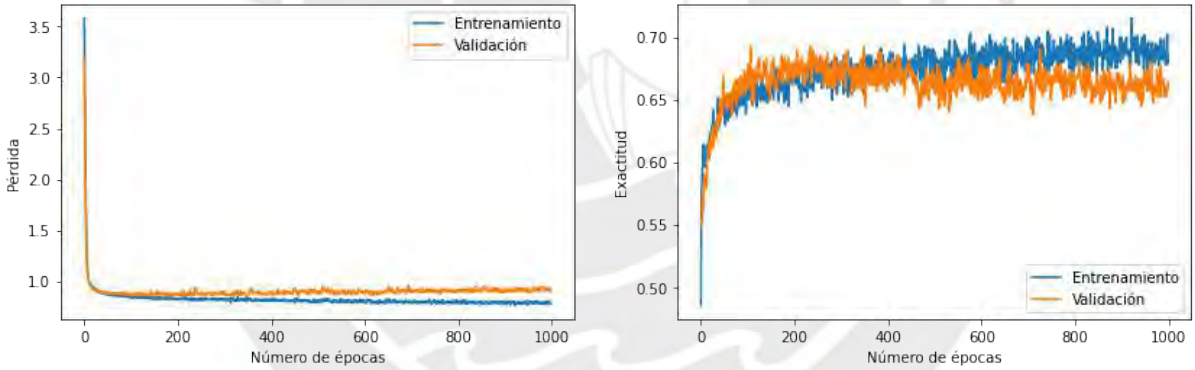
Por este motivo, se optó por emplear una capa de descarte con una tasa de 0.28 al inicializar el modelo y un regularizador L1 en la capa de entrada con un valor de 0.01. Se eligió al regularizador L1 por dos motivos. En primer lugar, debido a su capacidad de descarte de variables innecesarias, factor importante tomando en cuenta que la cantidad de variables independientes se elevó a 49 luego del proceso de *One-hot Encoding*. En segundo lugar, puesto que no se observó una mejora del desempeño al implementar un regularizador L2 o simultáneamente regularizadores L1 y L2. Como se puede apreciar en la figura 5.15, con el modelo definitivo se obtuvo un comportamiento más uniforme de los criterios de pérdida y exactitud para los conjuntos de entrenamiento y validación.



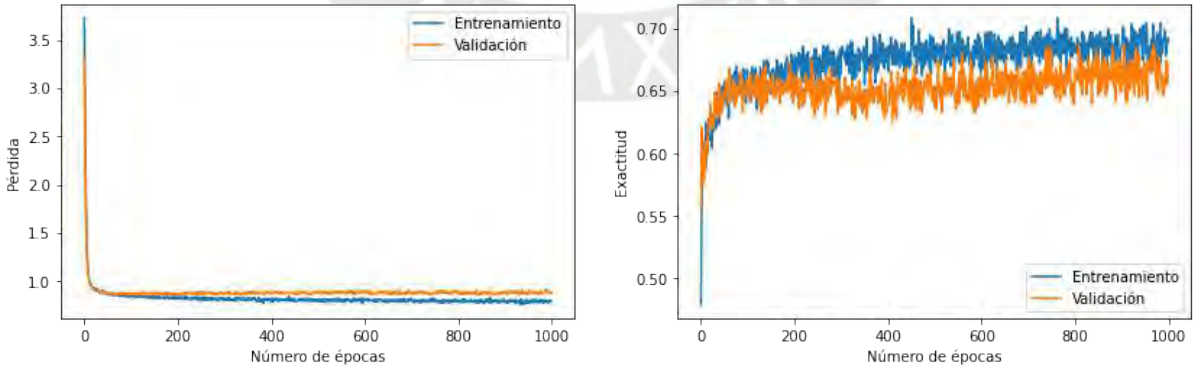
(a) Primer grupo



(b) Segundo grupo



(c) Tercer grupo



(d) Cuarto grupo

Figura 5.15. Variación de la pérdida y exactitud del modelo de redes neuronales sin segmentación geográfica con regularización L1 durante la validación cruzada.

Análisis del desempeño

Para este modelo, se obtuvo una exactitud de 0.69. Se presenta en la tabla 5.15 las métricas adicionales de precisión, exhaustividad y puntaje F1. Analizando los promedios ponderados de las métricas, se observa un puntaje F1 de 0.69, obtenido en base a valores de 0.69 tanto para precisión como para exhaustividad. Puesto que se obtuvieron valores idénticos de estas dos últimas métricas, se verifica que no se recae en la "paradoja de la exactitud". Cabe resaltar que al igual que en el caso del modelo de bosques aleatorios sin reducción de variables, los resultados más desfavorables se presentan respecto a la categoría intermedia B.

Tabla 5.15. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de redes neuronales sin segmentación geográfica y sin reducción de variables.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.67	0.81	0.73
B	0.54	0.43	0.48
A	0.81	0.75	0.78
Promedio ponderado	0.69	0.69	0.69

Las curvas ROC y el AUC obtenidos se presentan en la figura 5.16. Nuevamente, en esta gráfica se puede comprobar el comportamiento favorable del modelo para todas las categorías con excepción de la B, la cual tiene la menor área bajo la curva. Según el criterio del índice de Youden, los umbrales óptimos para las categorías A, B y C son de 0.47, 0.35 y 0.40 respectivamente. Las tasas de verdaderos positivos y falsos positivos correspondientes se detallan en la leyenda de la figura 5.16.

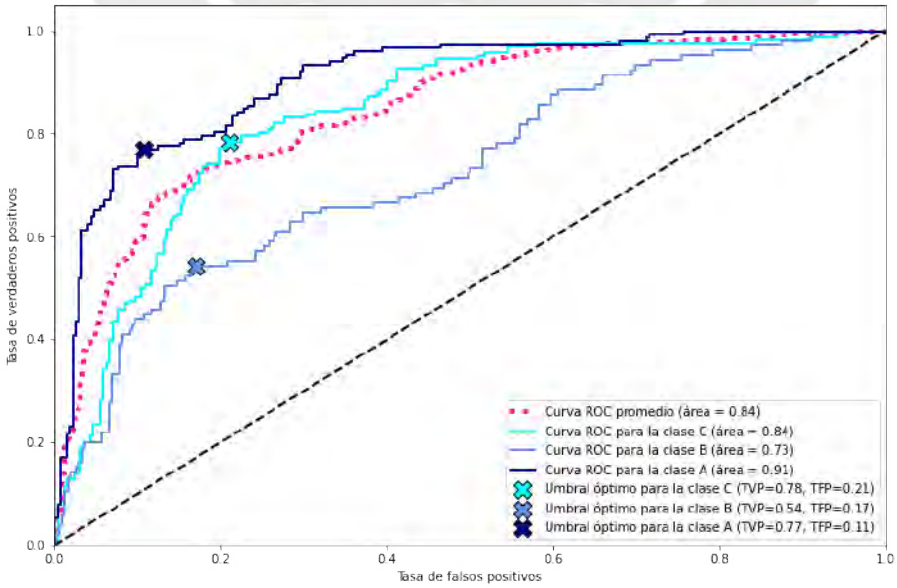


Figura 5.16. Curvas ROC correspondientes al modelo de redes neuronales sin segmentación geográfica y sin reducción de variables.

Análisis de importancia de variables

Las variables ordenadas por su importancia de acuerdo con el análisis de varianza se presentan en la figura 5.17. Para este primer modelo que considera la totalidad de variables independientes, se observa que las más importantes son: evapotranspiración ("EV" en la figura), distancia a fallas geológicas ("dfallas" en la figura) y precipitación promedio del año de evaluación ("precip_prom_anual" en la figura). Cabe resaltar que dos de estas variables, evapotranspiración y distancia a fallas, también se encontraron dentro de las cinco variables consideradas de alta importancia por su puntaje Relief-F. En lo concerniente a las variables categóricas, si bien estas obtuvieron una importancia considerablemente menor a las numéricas, las de mayor relevancia son: formación geológica del cuaternario pleistoceno-continental ("Qp-c" en la figura), relieve tipo valle, y formación geológica del cuaternario holoceno-continental ("Qh-c" en la figura). Respecto a las variables asociadas al relieve, la categoría de valle obtuvo una alta preponderancia ya que su morfología favorece la infiltración del agua a los acuíferos. En lo relacionado a las variables de formaciones geológicas, las formaciones "Qp-c" y "Qh-c" exhiben un potencial alto de agua subterránea dado que estos estratos recientes se componen principalmente por depósitos eólicos y aluviales que se componen principalmente por arenas de grano grueso o gravas gruesas redondeadas (Walsh Perú, 2009). Este tipo de suelo granular presenta cavidades de tamaño considerable en contraposición a suelos finos, por lo que su correlación a un alto potencial de agua subterránea es evidente.

En contraposición, las categorías asociadas al tipo de suelo obtuvieron una relevancia prácticamente nula en el modelo de redes neuronales. Idealmente, se esperaba que las variables de suelo arcilloso ("Cl" en la figura) y suelo arenoso sean las dominantes puesto que delimitan dos comportamientos extremos del suelo: el primero dificulta la percolación del agua y limita el potencial del agua subterránea por su porosidad, mientras que el segundo exhibe el comportamiento opuesto. Sin embargo, la variable de suelo arcilloso presentó una importancia prácticamente nula y la variable de suelo arenoso no figuró en el análisis de importancia. Esto se debe a que ambos parámetros no se encontraban adecuadamente representados en la base de datos; solo el 0.14 % de pozos presentaban un tipo de suelo arcilloso y ninguno presentaba un tipo de suelo formalmente clasificado como arenoso. En consecuencia, el modelo empleó a la categoría de suelo franco arcillo arenoso ("SaClLo" en la figura) para la clasificación ya que esta propiedad se encontraba presente en un 67.18 % de los pozos recopilados. Sin embargo, puesto que esta composición de suelo es sumamente heterogénea, no aporta información que el modelo pueda utilizar para diferenciar las categorías de potencial y, por ende, no contribuyó a mejorar las métricas de desempeño.

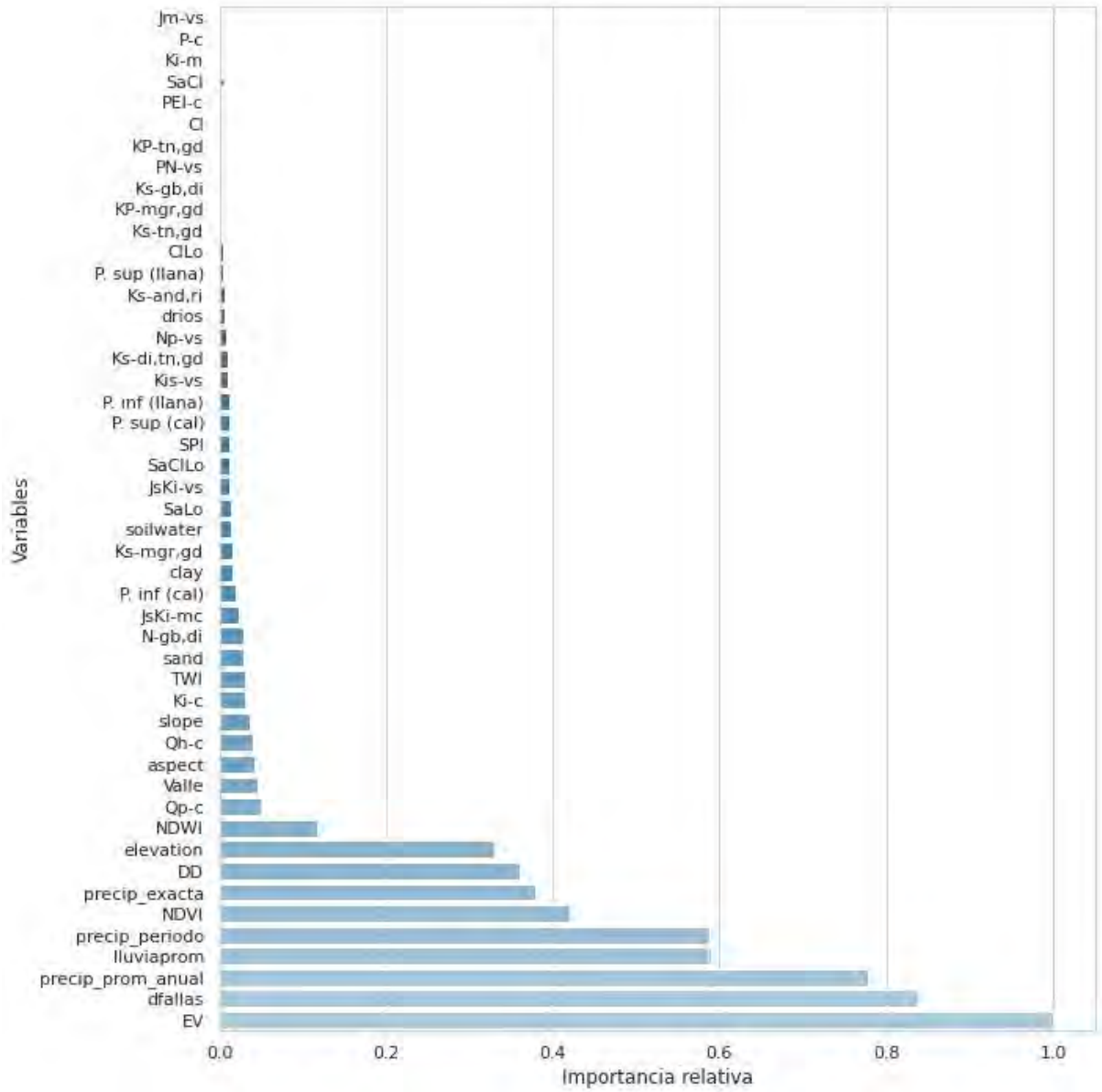


Figura 5.17. Análisis de importancia de variables correspondiente al modelo de redes neuronales sin segmentación geográfica y sin reducción de variables.

Reducción de variables

Análogamente al algoritmo de bosques aleatorios, se construyó un nuevo modelo descartando todas las variables de baja importancia y que según el análisis realizado previamente presentaran altos indicios de multicolinealidad según la matriz de correlación y el puntaje VIF. En este sentido, este nuevo modelo solo consideró las mismas 8 variables empleadas en el modelo reducido de bosques aleatorios: aspecto, densidad de drenaje, elevación, NDWI y las 4 variaciones de la variable precipitación. De manera similar al modelo de bosques aleatorios, se buscó determinar inicialmente cuál de las variaciones de la variable de precipitación es la que permite obtener un mejor desempeño en el problema de

clasificación. Sin embargo, como se puede observar en la figura 5.18, las métricas de desempeño más favorables se evidenciaron al usar las 4 variaciones en simultáneo.

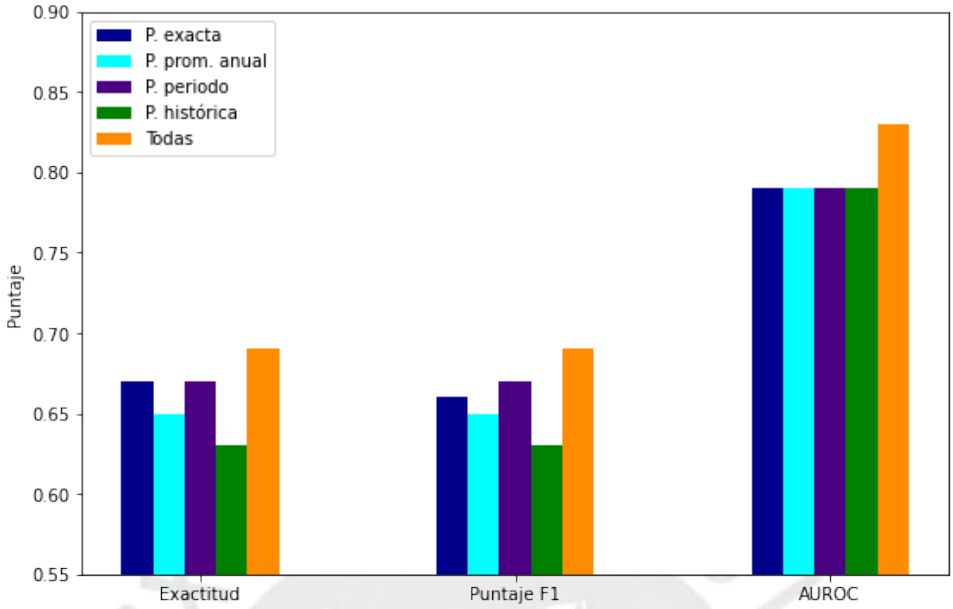


Figura 5.18. Análisis de importancia de los tipos de precipitación independientemente y en simultáneo para el modelo de redes neuronales.

Durante el proceso de validación cruzada, la arquitectura del modelo se optimizó incluyendo una capa oculta adicional de 50 neuronas. En la figura 5.19, se presenta la variación de las funciones de pérdida y exactitud conforme transcurrieron 1000 épocas de entrenamiento. Los puntajes ponderados obtenidos de precisión, exhaustividad y F1 fueron de 0.69 como se puede observar en la tabla 5.16. La exactitud obtenida análogamente fue de 0.69. El área bajo la curva ROC promedio presentada en la figura 5.20 es de 0.83, una centésima inferior al modelo anterior. Al haber prácticamente equiparado al modelo original a pesar de la reducción de 49 a 8 variables, se puede afirmar que la selección de variables fue adecuada.

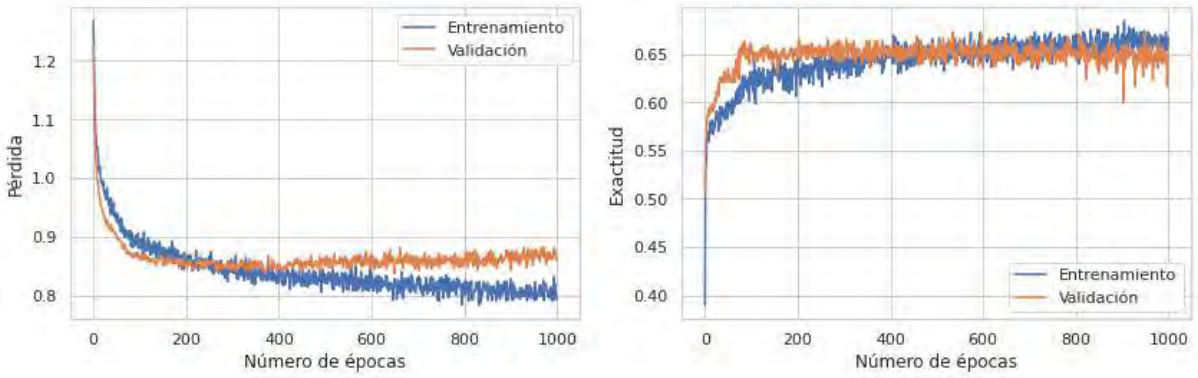


Figura 5.19. Variación de la pérdida y exactitud del modelo de redes neuronales con reducción de variables durante la validación cruzada.

Tabla 5.16. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de redes neuronales sin segmentación geográfica y con reducción de variables.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.70	0.81	0.75
B	0.53	0.48	0.50
A	0.80	0.70	0.75
Promedio ponderado	0.69	0.69	0.69

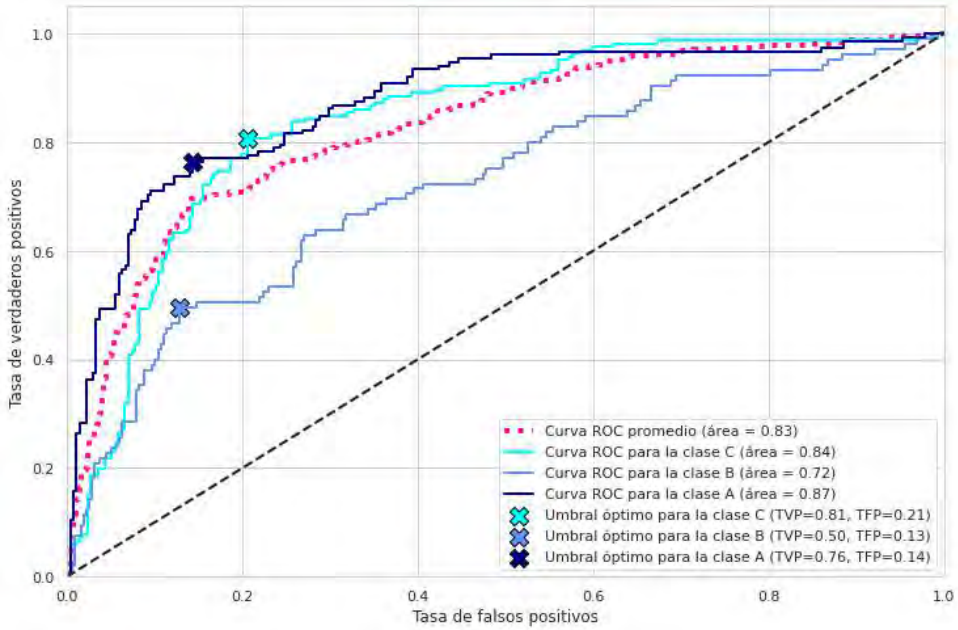


Figura 5.20. Curvas ROC correspondientes al modelo de redes neuronales sin segmentación geográfica y con reducción de variables.

El orden de importancia de variables se presenta en la figura 5.21. De manera similar al modelo de variables reducidas de bosques aleatorios, las propiedades de aspecto y NDWI fueron las menos relevantes para el desempeño del modelo. En contraste, para el modelo de redes neuronales las variables más importantes fueron densidad de drenaje, precipitación promedio histórica y precipitación promedio anual.

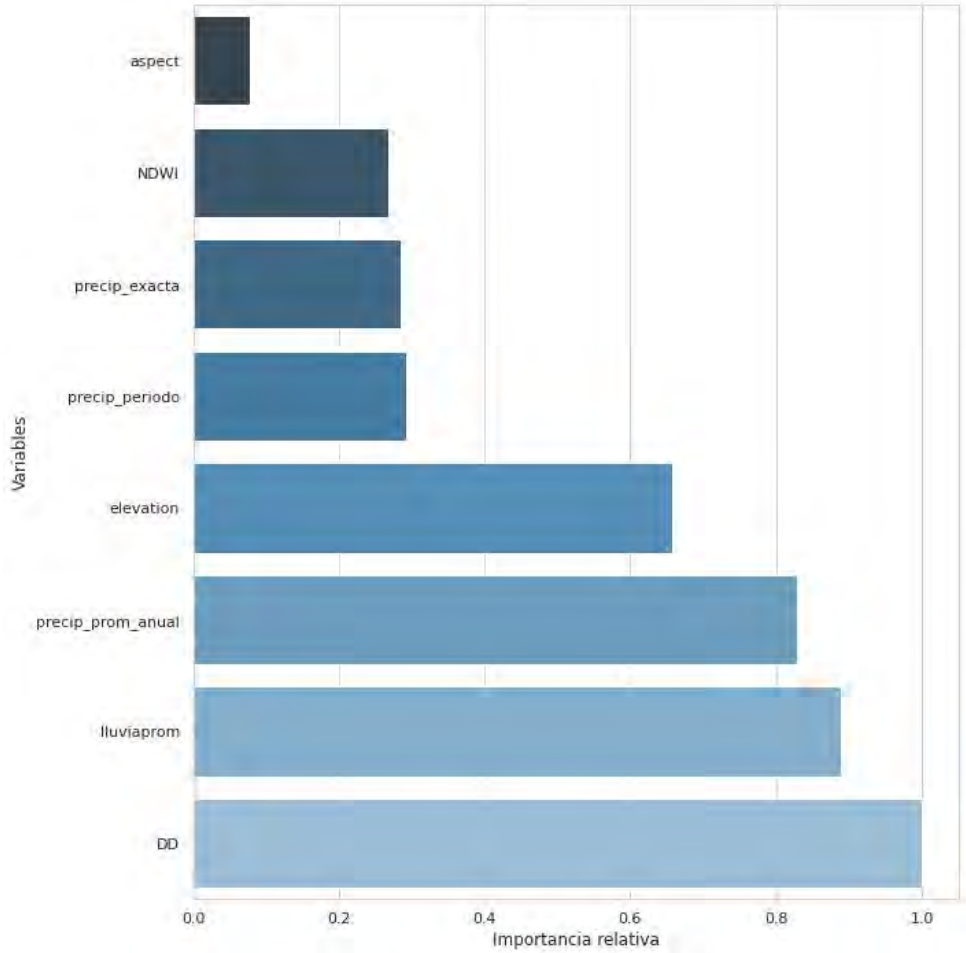


Figura 5.21. Análisis de importancia de variables correspondiente al modelo de redes neuronales sin segmentación geográfica y con reducción de variables.

5.3.2. Modelos con segmentación geográfica

Costa sur

Al concluir la etapa de validación cruzada, se maximizó el desempeño arquitectura consistente en dos capas ocultas de 25 neuronas cada una y regularizadores de descarte de 25 % y de tipo L1 en la capa de entrada. Se obtuvo una exactitud de 0.73 y un puntaje F1 ponderado de 0.70 como se puede apreciar en la tabla 5.17. Al igual que los modelos anteriores, se obtuvieron valores balanceados de precisión y exhaustividad y se observó un desempeño idóneo para todas las categorías a excepción de la intermedia.

Tabla 5.17. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de redes neuronales segmentado para la costa sur.

Clase	Precisión	Exhaustividad	Puntaje F1
C	0.68	0.94	0.79
B	0.69	0.30	0.41
A	0.85	0.72	0.78
Promedio ponderado	0.73	0.73	0.70

El área bajo la curva ROC promedio fue de 0.86, superior a la de los modelos sin segmentación. Los umbrales óptimos fueron de 0.42, 0.46 y 0.39 para las clases C, B y A respectivamente. Las tasas máximas de verdaderos positivos y mínimas de falsos negativos se presentan en la figura 5.22.

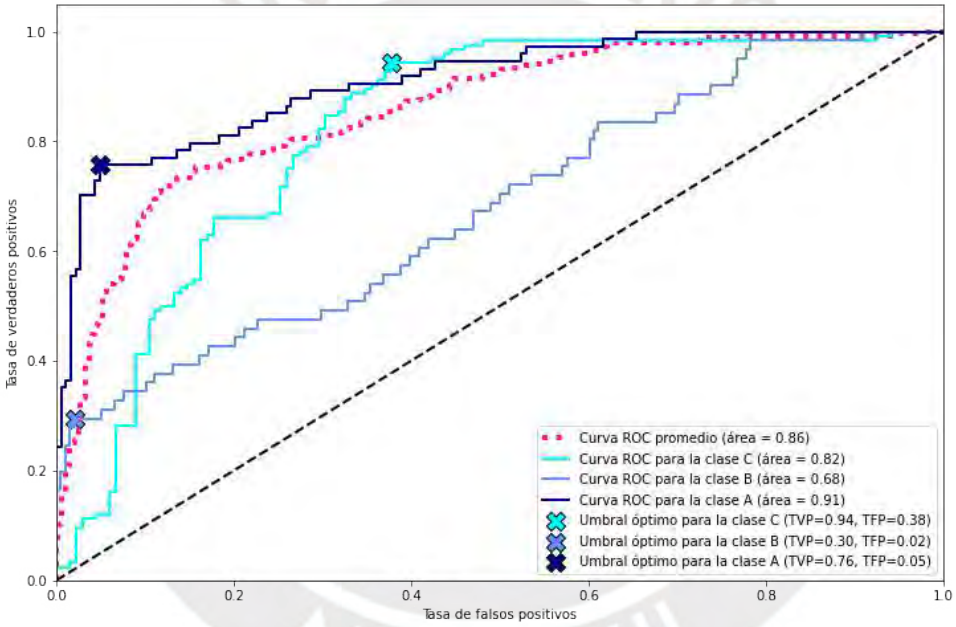


Figura 5.22. Curvas ROC correspondientes al modelo de redes neuronales segmentado para la costa sur.

Costa centro

Se obtuvo el mejor desempeño con una arquitectura de dos capas ocultas de 100 neuronas cada una y regularizadores de descarte de 15 % y de tipo L1 en la capa de entrada. Se obtuvo una exactitud de 0.72 y un puntaje F1 ponderado de 0.72 como se puede apreciar en la tabla 5.18.

Tabla 5.18. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de redes neuronales segmentado para la costa centro.

Clase	Precisión	Exhaustividad	Puntaje F1
B	0.76	0.82	0.79
A	0.64	0.55	0.59
Promedio ponderado	0.72	0.72	0.72

El área bajo la curva ROC promedio fue de 0.73, inferior a la de los modelos sin segmentación a diferencia del modelo de bosques aleatorios. El umbral óptimo determinado fue de 0.14. Las tasas máximas de verdaderos positivos y mínimas de falsos negativos se presentan en la figura 5.23.

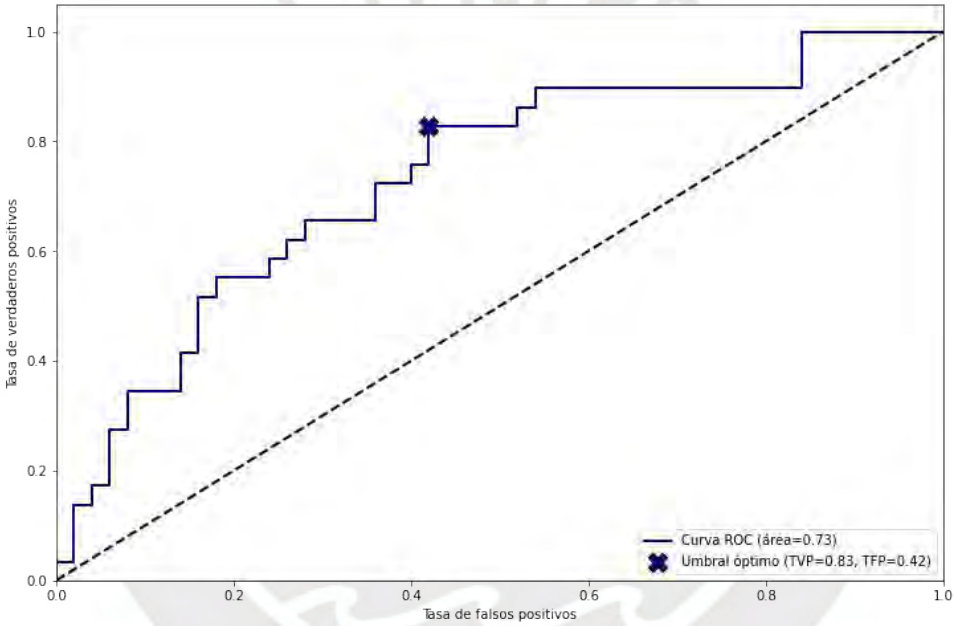


Figura 5.23. Curva ROC correspondiente al modelo de redes neuronales segmentado para la costa centro.

Costa norte

Al concluir la etapa de validación cruzada, se maximizó el desempeño con una arquitectura consistente en una capa oculta de 25 neuronas y regularizadores de descarte de 15 % y de tipo L1 en la capa de entrada. Se obtuvo una exactitud de 0.73 y un puntaje F1 ponderado de 0.73 como se puede apreciar en la tabla 5.19.

Tabla 5.19. Precisión, exhaustividad y puntaje F1 obtenidos para el modelo de redes neuronales segmentado para la costa norte.

Clase	Precisión	Exhaustividad	Puntaje F1
B	0.74	0.76	0.75
A	0.71	0.72	0.72
Promedio ponderado	0.73	0.73	0.73

El área bajo la curva ROC promedio fue de 0.71, nuevamente inferior a la de los modelos sin segmentación. El umbral óptimo determinado fue de 0.52. Las tasas máximas de verdaderos positivos y mínimas de falsos negativos se presentan en la figura 5.24.

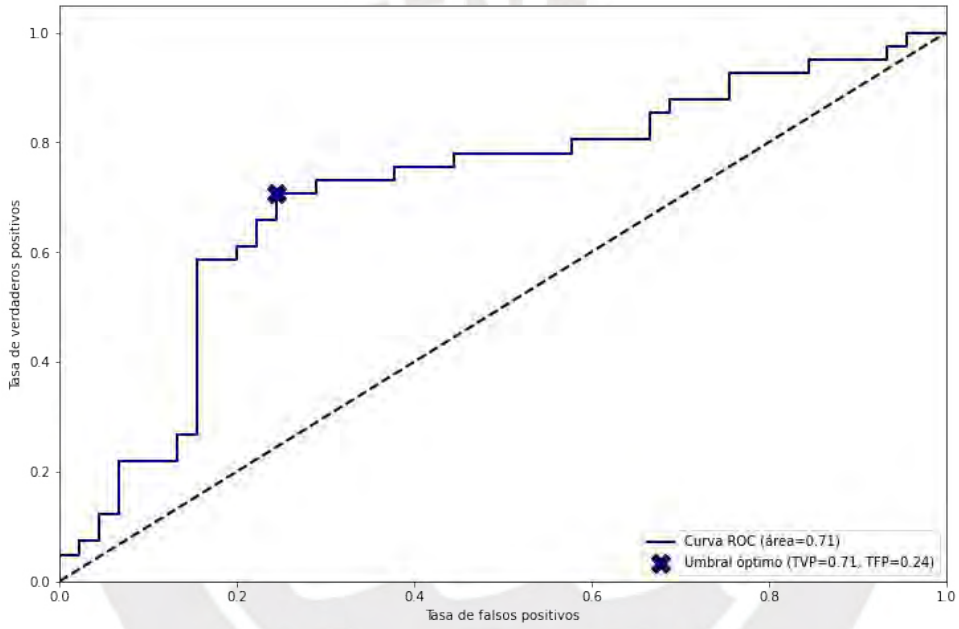


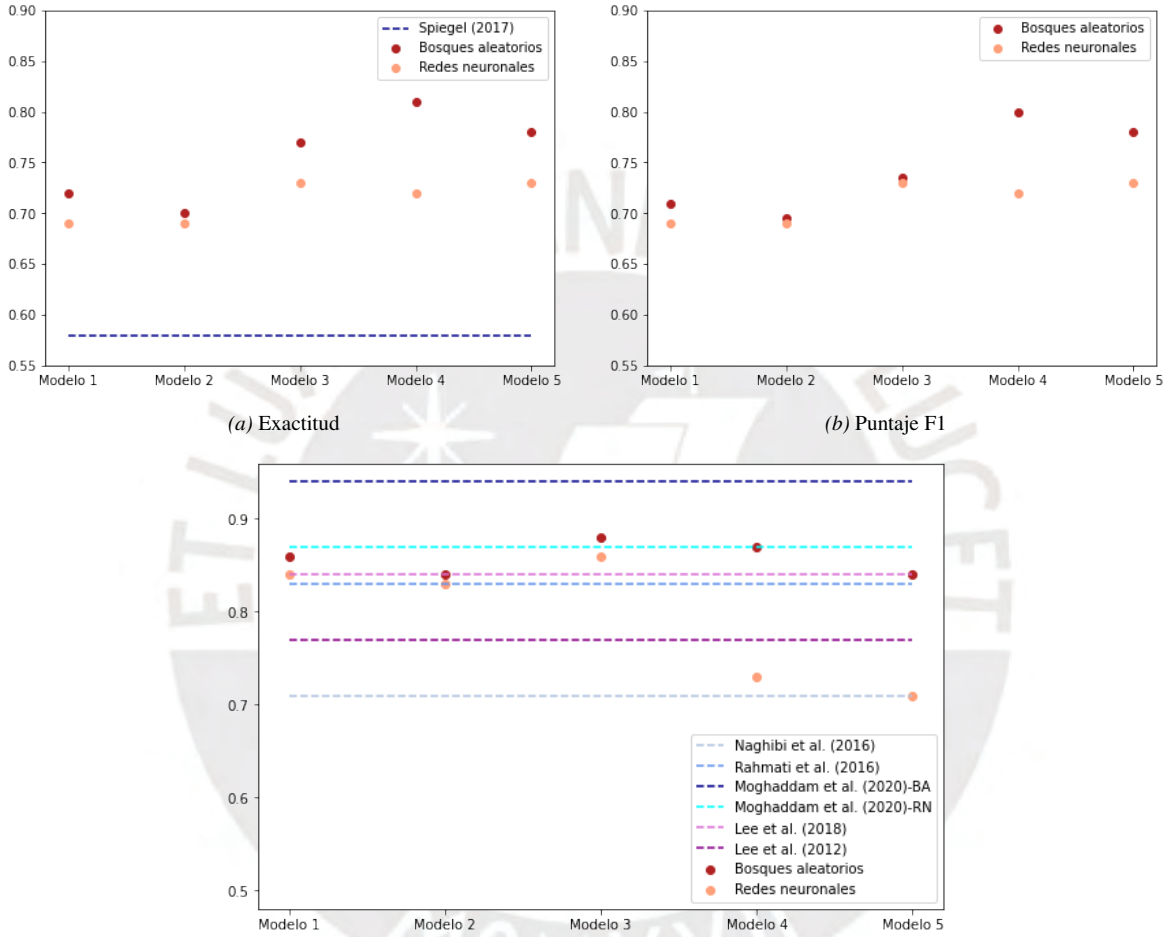
Figura 5.24. Curva ROC correspondiente al modelo de redes neuronales segmentado para la costa norte.

5.4. Generalización de los modelos de clasificación

5.4.1. Comparación de modelos

Con el objetivo de determinar cuál algoritmo permitió construir el modelo con mejor desempeño, se resumió en la figura 5.25 los resultados obtenidos para las tres métricas consideradas en esta investigación: exactitud, puntaje F1 ponderado y AUC. Tanto para el caso de bosques aleatorios como para el de redes neuronales se consideraron cinco modelos: sin segmentación geográfica y considerando la totalidad de variables (modelo 1 en la figura), sin

segmentación geográfica pero solo considerando las 8 variables más importantes (modelo 2 en la figura), con segmentación aplicada a la costa norte (modelo 3 en la figura), costa centro (modelo 4 en la figura) y costa sur (modelo 5 en la figura). Para propósitos comparativos, en las figuras también se incluyen las métricas de desempeño obtenidas para los modelos de bosques aleatorios propuestos por Spiegel (2017), Naghibi et al. (2016), Rahmati et al. (2016) y Moghaddam et al. (2020). Asimismo, se incluyen las métricas correspondientes a los modelos de redes neuronales de Lee et al. (2018, 2012) y Moghaddam et al. (2020).



(a) Exactitud

(b) Puntaje F1

(c) Área bajo la curva ROC

Figura 5.25. Comparación de las tres métricas de evaluación obtenidas para los modelos planteados en la presente investigación y los del estado del arte.

A partir de estas gráficas, se puede observar que los cinco modelos construidos empleando el algoritmo de bosques aleatorios presentan mejores métricas de desempeño de manera unánime en comparación a los generados con el algoritmo de redes neuronales. Estos resultados concuerdan con los obtenidos por Moghaddam et al. (2020) para la cuenca de Hableh-Roud. En el mapeo de potencial de agua subterránea, la superioridad de los modelos de bosques aleatorios se explica debido a la cantidad limitada de datos (pozos con caudal y coordenadas conocidas) para el entrenamiento, validación y evaluación de los modelos. Si

bien los modelos de redes neuronales se basan en algoritmos computacionalmente más complejos, requieren una disponibilidad considerablemente mayor de datos para mejorar su desempeño. Evidencia de este comportamiento se puede apreciar en el análisis de importancia de variables efectuado en la presente investigación para ambos algoritmos en las figuras 5.7 y 5.17. El modelo de bosques aleatorios es capaz de descartar rápidamente las variables menos importantes asignándoles una relevancia nula para mejorar el desempeño. Sin embargo, en el caso de las redes neuronales, la distribución de importancia se realiza de manera más uniforme, impidiendo jerarquizar de manera adecuada múltiples variables pese a las múltiples épocas de entrenamiento. En este sentido, se estima que, de disponer información de caudal de los 52 930 pozos registrados en la base de datos de la ANA, se podrían obtener modelos de redes neuronales con un mejor desempeño. No obstante, un aspecto positivo que cabe resaltar de los modelos construidos empleando redes neuronales es su capacidad de aprovechar de mejor manera las variables categóricas. Si bien estas son descartadas inmediatamente por el modelo de bosques aleatorios, el modelo de bosques aleatorios fue capaz de identificar formaciones geológicas favorables al potencial de agua subterránea.

Por otro lado, comparando los resultados obtenidos con el estado del arte, se puede resaltar que los cinco modelos obtenidos para ambos algoritmos superan en términos de exactitud al modelo propuesto por Spiegel (2017). En lo que respecta al parámetro AUC, con excepción del modelo de redes neuronales segmentado geográficamente para la costa norte y costa centro, los modelos superan a los propuestos por Lee et al. (2012), Naghibi et al. (2016) y Rahmati et al. (2016). Sin embargo, únicamente el modelo de bosques aleatorios segmentado geográficamente para la costa sur presenta un desempeño superior al de redes neuronales de Moghaddam et al. (2020). Además, cabe resaltar que el modelo de bosques aleatorios de este autor, supera a todos los planteados en la presente investigación. Una de las razones principales por las cuáles se obtuvieron resultados inferiores se explica por la disponibilidad de información. En el modelo de Moghaddam et al. (2020) se computan modelos predictivos para caracterizar el potencial de agua subterránea asociado a la cuenca de Hableh-Roud. Para este proceso, se disponen 7.3 pozos/km^2. En cambio, para el caso de los modelos sin segmentación en la presente investigación se busca caracterizar la totalidad de la costa con 2127 pozos, es decir, una tasa de 0.01 pozos/km^2, sumamente inferior. Asimismo, considerando todo el territorio nacional e incluyendo en los cálculos los pozos de la sierra y de la selva con caudal registrado, la tasa se reduce aún más a 0.002 pozos/km^2. De disponer la información de caudal para la totalidad de pozos en el Perú como se mencionó previamente, la tasa sería de 0.04 pozos/km^2, 20 veces la que se dispone actualmente pero aun considerablemente inferior a la del estado del arte.

Para analizar el efecto de la reducción de variables en el desempeño, se compararon los modelos 1 y 2. Se puede apreciar que la influencia de la reducción de variables de 49 a 8 es

despreciable. En el caso de las variables numéricas, esto se debe a la alta multicolinealidad que existe entre las variables seleccionadas, las cuales afectaron la interpretabilidad de la importancia de variables en los modelos de ambos algoritmos. Por ejemplo, en los modelos que consideraban a la totalidad de variables se identificaba a la evapotranspiración como al factor más relevante en la estimación del potencial de agua subterránea. Sin embargo, dado que este factor puede ser explicado linealmente por el resto de las variables, se puede descartar sin afectar el desempeño. Esto demuestra que, para el área de estudio, no se requiere más que 8 variables para una clasificación adecuada del potencial de agua subterránea: aspecto, densidad de drenaje, elevación, NDWI y las 4 variaciones de la variable precipitación. Este resultado es similar al obtenido por Naghibi et al. (2016) y Rahmati et al. (2016), ya que en ambas investigaciones se identificó a la elevación y a la densidad de drenaje como las variables más importantes para la clasificación del potencial de agua subterránea. No obstante, Moghaddam et al. (2020) identificó a la posición relativa de la pendiente y a la litología como los factores más relevantes. En ninguno de los tres se emplearon variables de precipitación para la construcción de los modelos predictivos.

Finalmente, se procedió a analizar el impacto de la segmentación en los modelos. Asesorando al modelo correspondiente a la costa sur (modelo 3), se puede apreciar un incremento de todas las métricas utilizadas para cuantificar el desempeño de los modelos: exactitud, puntaje f1 y área bajo la curva ROC. Esto se debe a que la mayoría de los pozos disponibles corresponden a esta zona y al reducir el área geográfica de estudio, se dispone de una tasa de 0.02 pozos/km^2, el doble en comparación a los modelos no segmentados. En el caso de la costa centro y norte (modelos 4 y 5), si bien se evidencia una mejora del desempeño en cuestiones de exactitud y puntaje F1, se aprecia una disminución del área bajo la curva ROC para los modelos de redes neuronales. La reducción de esta métrica se debe a que a que la minoría de los pozos de caudal conocido se encuentran en estas regiones, reduciendo la disponibilidad de datos a 0.008 pozos/km^2 en el caso de la costa norte. Este factor repercute más severamente en los modelos de redes neuronales debido a su exigencia numérica de gran cantidad de datos para el entrenamiento y validación.

5.4.2. Extrapolación a otras regiones del modelo con mejor desempeño

Puesto que el modelo de bosques aleatorios presentó un mejor desempeño que el de redes neuronales en la presente investigación, se procedió a emplearlo para la clasificación de la totalidad del territorio nacional de acuerdo con su potencial de agua subterránea. Previo a ello, se procedió a verificar el desempeño del modelo con pozos ubicados en regiones fuera de la costa. A partir de los datos disponibles, se disponen pozos de caudal conocido únicamente en las localidades de Puno e Iquitos. El modelo acertó un 83.3 % del total de pozos disponibles en

Punto y 100 % de los pozos de la región de Iquitos. Si bien estos resultados favorecen la aplicabilidad del modelo para las regiones sierra y selva respectivamente, resulta imprescindible disponer de más datos de validación en estas regiones para afirmar que el modelo es totalmente aplicable a nivel nacional. En todo caso, habiendo obtenido resultados favorables con los datos disponibles, se procedió a aplicar el modelo en el territorio peruano considerando píxeles de 1000x1000m para la clasificación. Para ello, se recreó la arquitectura del modelo sin segmentación geográfica de bosques aleatorios empleando la adaptación de la librería *Smile* de Java, disponible en GEE. El resultado de este proceso se presenta en la figura 5.26.

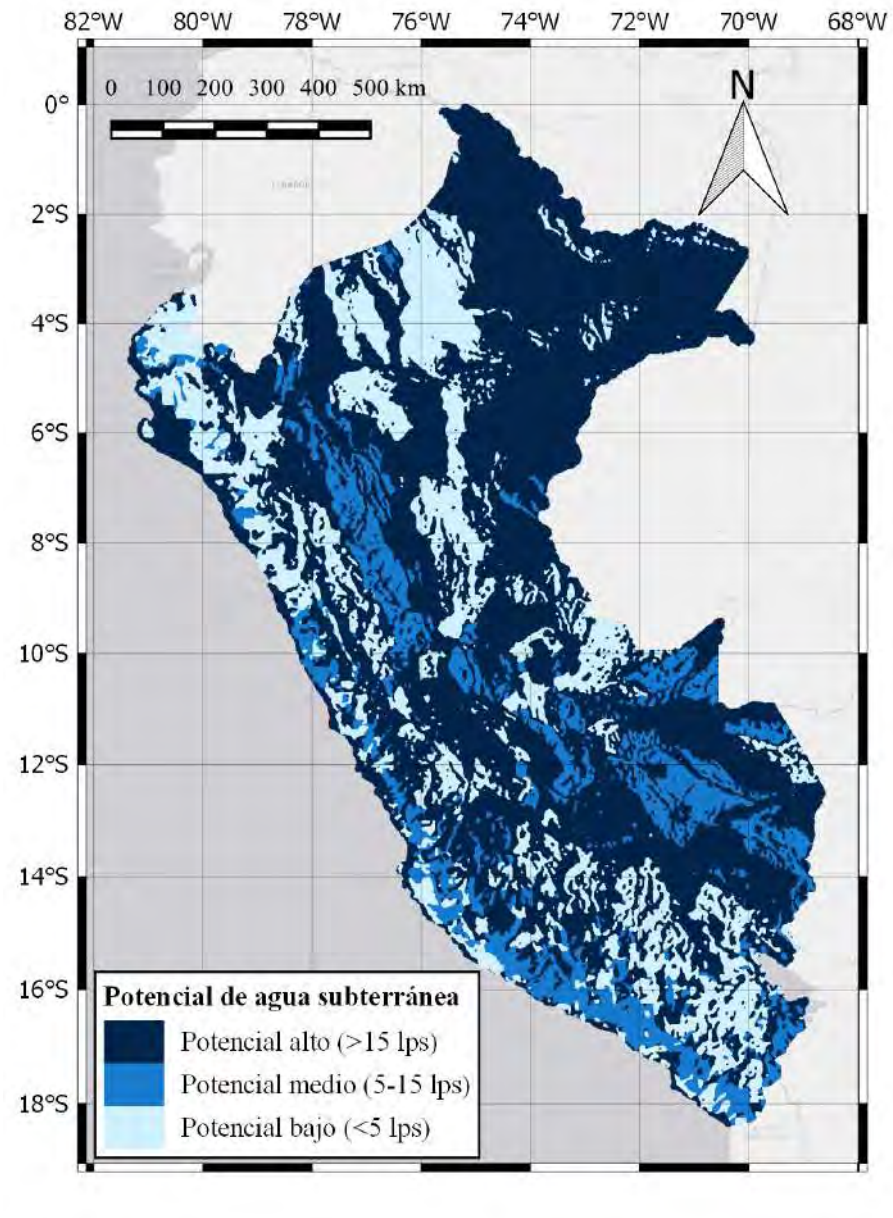
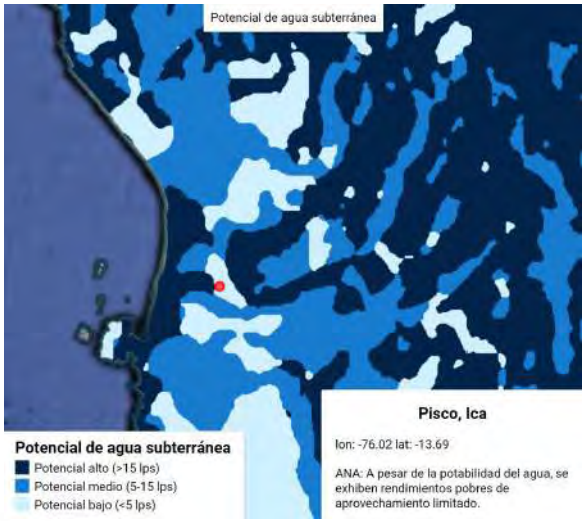


Figura 5.26. Clasificación del potencial de agua subterránea en el territorio peruano.

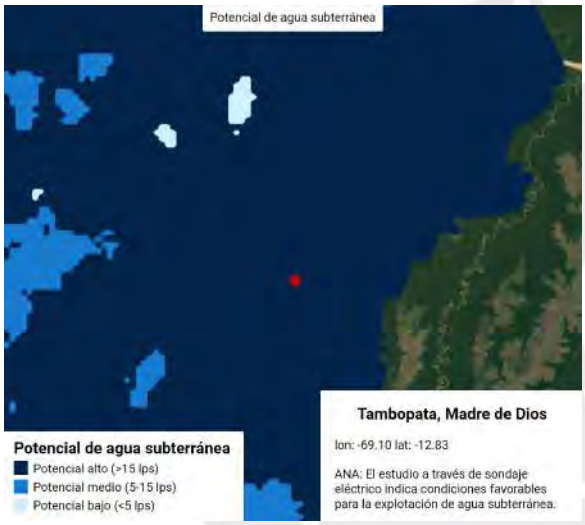
Cabe resaltar que se empleó un remuestreo de la imagen original empleando una operación focal de moda para cada píxel considerando un kernel circular de 3 píxeles de radio. Gracias a este procedimiento, se pudieron obtener regiones claramente diferenciadas de potencial de agua subterránea y suavizar las transiciones entre ellas. Analizando el resultado final obtenido, se tiene que un 59.8 % del territorio del Perú se encuentra en la categoría C, mientras que las categorías B y A suponen valores de 16.1 % y 24.2 % respectivamente. En otras palabras, se estima que 311022.20 km^2 del Perú posee un alto potencial de agua subterránea. De todo el territorio con alto potencial, la región selva es la que contiene la mayor parte con un 42.8 % clasificado como de tipo A.

Una validación de este mapa se realizó comparándolo con los resultados obtenidos en estudios hidrogeológicos a nivel cuenca efectuados por la ANA. Para ello se creó una aplicación de visualización interactiva que puede ser accedida en el enlace https://cesarportocarrero.users.earthengine.app/view/gwp-peru (versión para computadoras) o en https://cesarportocarrero.users.earthengine.app/view/gwp-peru-movil (versión para dispositivos móviles). Como se puede apreciar, los resultados coinciden con los análisis in situ de las cuencas de Sama, Pisco, Tambopata, Santa, Ventanilla, Jequetepque, Huipoca, San Jerónimo y Salcedo (figuras 5.27a a 5.27i). No obstante, erra en el caso de Motupe (figura 5.27j) debido a la poca cantidad de pozos en la costa norte. Asimismo, la ausencia total de pozos impide que delimite algunas zonas conocidas por su alto potencial como Matarani en Arequipa. Por otra parte, al presentar una gran mayoría de pozos en la costa, variables críticas para el condicionamiento de pozos en otras regiones son menospreciadas por los modelos. Tal es el caso del tipo de suelo arcilloso en la selva, el cual impide caudales altos de extracción en esta región a pesar de las grandes reservas subterráneas existentes. Al no existir pozos con esta característica en la costa, el modelo es incapaz de aprender esta tendencia y clasifica a gran parte de la selva como de alto potencial.

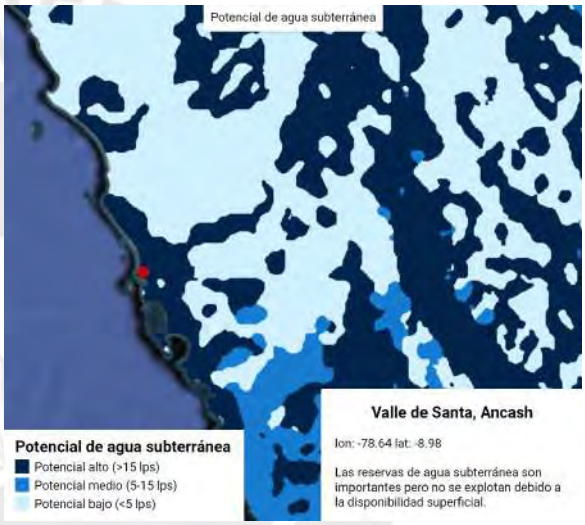
(a) Sama



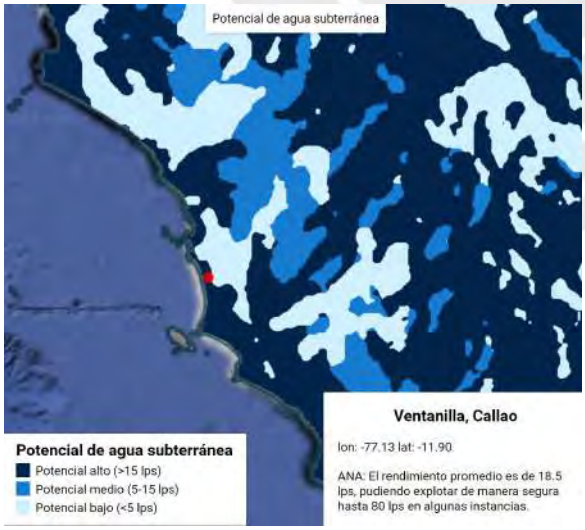
(b) Pisco



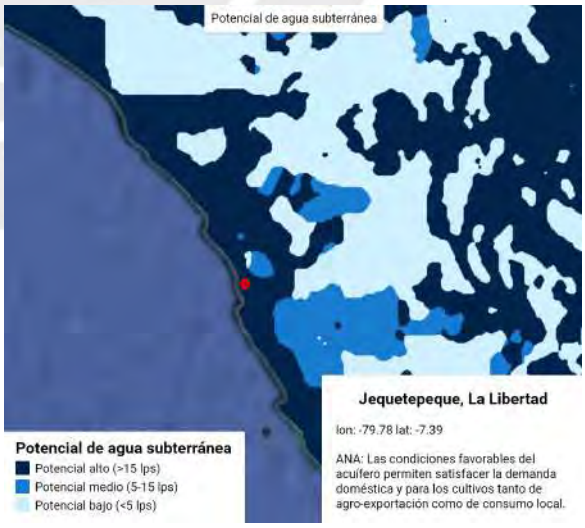
(c) Tambopata



(d) Santa

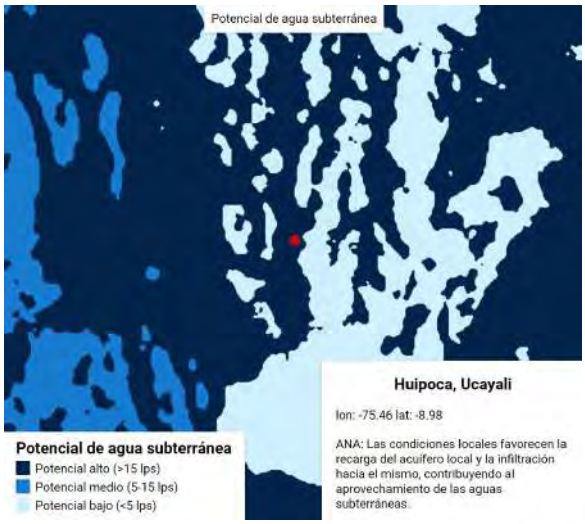


(e) Ventanilla

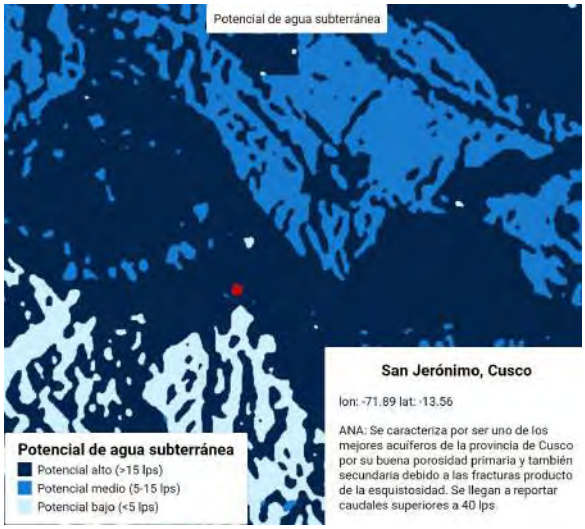


(f) Jequetepeque

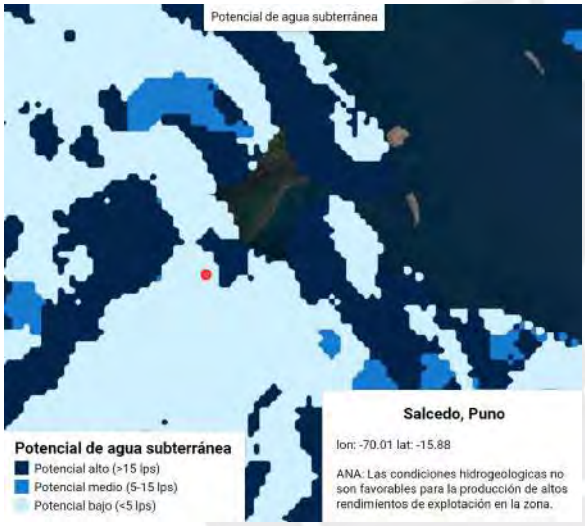
Figura 5.27. Comparación entre el potencial de agua subterráneo reportado por el ANA mediante estudios in situ y el resultado del modelo de bosques aleatorios construido.



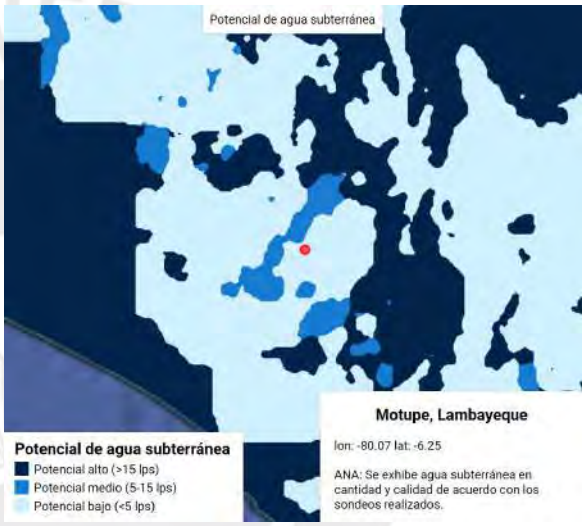
(g) Huipoca



(h) San Jerónimo



(i) Salcedo



(j) Motupe

Figura 5.27 (cont.). Comparación entre el potencial de agua subterráneo reportado por el ANA mediante estudios in situ y el resultado del modelo de bosques aleatorios construido.

Capítulo VI

Conclusiones

La presente investigación ha permitido explorar las posibilidades de la integración simultánea de tecnologías del estado del arte de ciencia de datos, telemedición y sistemas de información geográfica al ámbito de la hidrología subterránea. Los dos modelos de clasificación de aprendizaje de máquina han obtenido indicadores de desempeño en términos de métricas de exactitud, puntaje F1 y AUC equiparables a los del estado del arte en Irán, Zambia y Corea del Sur. Una validación más orientada a los tomadores de decisión fue realizada contemplando los estudios hidrogeológicos in situ efectuados por la ANA. Resultados no favorables fueron identificados únicamente en lugares con total ausencia de pozos de caudal conocido. En este sentido, se puede afirmar que los modelos de aprendizaje de máquina sí son útiles para la clasificación del potencial de agua subterránea siempre y cuando se disponga un inventario adecuadamente distribuido de pozos para el entrenamiento de los modelos. Al ser el presente el primer trabajo de esta índole en el Perú, se espera que el marco metodológico propuesto sirva como base para el desarrollo de modelos más potentes. El código se encuentra disponible en https://github.com/cesport/Tesis.

En lo que concierne al análisis de importancia de variables en la categorización del potencial del agua subterránea, se puede concluir que son ocho las variables de mayor preponderancia: aspecto topográfico, densidad de drenaje, elevación, NDWI, precipitación en un periodo de más y menos cinco años, precipitación exacta, precipitación anual y precipitación histórica. Se validó la elección al confirmar que el desempeño de los modelos considerando la totalidad de variables era equiparable al que solo consideraba ocho. En contraste a la hipótesis planteada inicialmente, las variables de evapotranspiración y tipo de suelo no fueron vitales para el desempeño de los modelos. En el caso de la evapotranspiración, esta se descartó debido a que esta propiedad puede ser explicada a partir de una combinación lineal del resto de variables, por lo que es redundante y ocasiona multicolinealidad. Por otro lado, en el caso del tipo de suelo, esta variable no fue identificada como relevante debido a que los pozos disponibles para el entrenamiento se ubicaban predominantemente en la costa.

Teniendo en cuenta que en esta región el tipo de suelo es prioritariamente arenoso, fue imposible para los algoritmos aprender como los suelos cohesivos son perjudiciales para los caudales de explotación. Esta parcialización hacia la costa del análisis de importancia de variables explica también la importancia otorgada a las variables de precipitación. En ausencia de variaciones considerables de topografía, geología o tipo de suelo en los datos de entrenamiento, estas variables no contribuyeron al desempeño de los modelos, a pesar de que pudiesen ser más importantes para el análisis del caudal subterráneo en la sierra y en la selva.

Respecto al desempeño de los dos algoritmos planteados para el caso de estudio, se comprobó que los modelos basados en bosques aleatorios son superiores a los basados en redes neuronales para el presente caso de estudio de acuerdo con las tres métricas propuestas. Esto se debe a que los bosques aleatorios tienden a aprender a reconocer patrones de mejor manera en un contexto de escasez de información. En contraposición, la complejidad computacional de las redes neuronales las favorece en caso exista una abundancia de información para el entrenamiento. Dado que en la presente investigación el número de pozos adecuados para la construcción de modelos era sumamente inferior a la del estado del arte (aproximadamente 0.55 % de los inventarios empleados por otros investigadores) y se encontraban irregularmente distribuidos en el territorio nacional, el desempeño de los modelos de redes neuronales se vio perjudicado. Asimismo, este factor afectó a la influencia de la segmentación geográfica en el desempeño de los modelos. Utilizar únicamente pozos de la costa sur permitió disminuir la varianza de las variables, mejorando la capacidad de aprendizaje del modelo, lo cual se reflejó en el incremento de las métricas de evaluación. No obstante, esto solo fue posible dado que esta región concentra la mayoría de los pozos de caudal conocido inventariados por la ANA. En el caso de la costa centro y norte, la disminución de la disponibilidad de información al segmentar el área de estudio impide mejorar las métricas de desempeño manteniendo una clasificación de tres categorías.

Para futuros trabajos relacionados al tema, se recomienda delimitar zonas de amortiguamiento dentro de los radios de influencia de los pozos en estado de explotación. En estas zonas se debe evitar la construcción de pozos debido a la depresión de la napa freática dentro del emplazamiento. De esta manera, se podría incluir el impacto humano en el potencial de agua subterránea. Asimismo, se sugiere utilizar los algoritmos de aprendizaje automatizado para clasificar la calidad del agua subterránea. Si bien el modelo desarrollado en esta investigación califica el volumen de extracción, no explora el estado de esta. Empleando los valores de dureza, pH y conductividad eléctrica de la base de datos del ANA se podría evaluar la portabilidad de agua para consumo doméstico y la calidad de agua para riego según la escala de Wilcox. Estas propuestas permitirían aportar más información sobre el estado del reservorio subterráneo, incrementando el valor de la herramienta para los tomadores de decisión durante la fase de anteproyecto de construcción de pozos.

Bibliografía

Adiat, K. A., Nawawi, M. N., y Abdullah, K. (2012). Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool - A case of predicting potential zones of sustainable groundwater resources. *Journal of Hydrology*, *440-441*, 75–89. doi: 10.1016/j.jhydrol.2012.03.028

AGI. (2018). *How to find groundwater using Electrical Resistivity Imaging*. Descargado 2019-05-19, de https://bit.ly/2LW1MS8

Akinlalu, A., Adegbuyiro, A., Adiat, K., Akeredolu, B., y Lateef, W. (2017). Application of multi-criteria decision analysis in prediction of groundwater resources potential: A case of Oke-Ana, Ilesa Area Southwestern, Nigeria. *NRIAG Journal of Astronomy and Geophysics*, *6*(1), 184–200. doi: 10.1016/j.nrjag.2017.03.001

Allen, M. P. (2007). The problem of multicollinearity. En *Understanding regression analysis* (pp. 176–180). Boston: Springer US. doi: 10.1007/978-0-585-25657-3_37

ANA. (2012). *Perú: El agua en cifras.* Descargado 2019-08-29, de https://bit.ly/2phXE4M

ANA. (2016). *Ministerio de Agricultura y Riego ejecuta plan de contingencia por peligro inminente de déficit hídrico en el sur.* Descargado 2020-06-19, de https://www.ana.gob.pe/noticia/ministerio-de-agricultura-y-riego-ejecuta-plan-de-contingencia-por-peligro-inminente-de

Belletich, E. (2015). *Poechos: el coloso del Perú.* Descargado 2020-06-19, de http://udep.edu.pe/hoy/2015/poechos-el-coloso-del-peru/

Bennie, J., Hill, M. O., Baxter, R., y Huntley, B. (2006). Influence of slope and aspect on long-term vegetation change in British chalk grasslands. *Journal of Ecology*, *94*(2), 355–368. doi: 10.1111/j.1365-2745.2006.01104.x

Bense, V. F., Gleeson, T., Loveless, S. E., Bour, O., y Scibek, J. (2013). Fault zone hydrogeology. *Earth-Science Reviews*, *127*, 171–192. Descargado de https://bit.ly/2MpBZjY doi: 10.1016/j.earscirev.2013.09.008

Bhanja, S. N., Malakar, P., Mukherjee, A., Rodell, M., Mitra, P., y Sarkar, S. (2019). Using Satellite-Based Vegetation Cover as Indicator of Groundwater Storage in Natural Vegetation Areas. *Geophysical Research Letters*. doi: 10.1029/2019gl083015

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–

231. doi: 10.1214/SS/1009213726

Breiman, L., Friedman, J., Stone, C. J., y Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis. Descargado de https://books.google.com.pe/books?id=JwQx-WOmSyQC

Breiman, L., Friedman, J. H., Olshen, R. A., y Stone, C. J. (2017). *Classification and regression trees*. CRC Press. doi: 10.1201/9781315139470

Bustos, X., y Bermúdez, M. (2017). Determinación y comparación de índices de erosión teóricos en cuencas del flanco surandino venezolano, apoyado en sistemas de información geográfica y programación Python. *Terra nueva etapa*, *33*(53), 105–122. Descargado de https://bit.ly/2Ixxh2H

C3S. (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.*

Campillo, G. (2010). *Caudal critico en pozos profundos-Temas de interés-Aguamarket.* Descargado 2019-05-29, de https://bit.ly/31U5UYc

Castañeda, M., y Céspedes, J. (2017). *Registro de pozos.* Descargado 2020-03-31, de http://snirh.ana.gob.pe/visorPozos/

Chander, G., Markham, B. L., y Helder, D. L. (2009). Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sensing of Environment*, *113*(5), 893–903. doi: 10.1016/j.rse.2009.01.007

Chow, V. T. (1988). *Applied Hydrology* (1a ed.). New York: McGraw-Hill.

Chowdhury, A., Jha, M. K., Chowdary, V. M., y Mal, B. C. (2008). Integrated remote sensing and GIS-based approach for assessing groundwater potential in West Medinipur district, West Bengal, India. *International Journal of Remote Sensing*, *30*(1), 231–250. doi: 10.1080/01431160802270131

Conagua. (2015). *Agua Subterránea.* Descargado 2019-05-26, de https://bit.ly/2n8L3RY

Craney, T. A., y Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403. Descargado de http://www.tandfonline.com/doi/abs/10.1081/QEN-120001878 doi: 10.1081/QEN-120001878

Didan, K. (2015). *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006.* doi: https://doi.org/10.5067/MODIS/MOD13Q1.006

Dijon, R. (1981). *Groundwater exploration in crystalline rocks in Africa.* Descargado 2019-06-24, de https://bit.ly/2Vlrq5K

Driscoll, F. (1987). *Groundwater and Wells,* (2.ª ed.). Minnesota: Johnson Division.

Emanuel, C., y Escurra, J. (2000). *Manejo integrado de los recursos hídricos.* Descargado 2019-05-26, de https://bit.ly/2LWuJNN

Falah, F., Ghorbani Nejad, S., Rahmati, O., Daneshfar, M., y Zeinivand, H. (2017). Applicability of generalized additive model in groundwater potential modelling and comparison its performance by bivariate statistical methods. *Geocarto International*, *32*(10), 1069–1089. doi: 10.1080/10106049.2016.1188166

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., ... Alsdorf, D. (2007).

The Shuttle Radar Topography Mission. *Reviews of Geophysics*, *45*(2). Descargado de http://doi.wiley.com/10.1029/2005RG000183 doi: 10.1029/2005RG000183

Folch, A. (s.f.). *Geological and human influences on groundwater flow systems in range-and-basin areas: the case of the Selva Basin (Catalonia, NE Spain)* (Tesis Doctoral no publicada). Universidad Autónoma de Barcelona,.

Gestmontes. (2010). *¿Cuánto cuesta hacer un Pozo?* Descargado 2019-05-19, de https://bit.ly/35fJ7Za

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., y Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. doi: 10.1016/j.rse.2017.06.031

Guevara, A. (2014). *La crisis del agua en Perú.* Descargado 2019-08-29, de https://bit.ly/339Ow1U

Guresen, E., y Kayakutlu, G. (2011). Definition of Artificial Neural Networks with comparison to other networks. En *Procedia computer science* (Vol. 3, pp. 426–433). Istanbul: Elsevier. doi: 10.1016/j.procs.2010.12.071

Hartmann, A., Gleeson, T., Wada, Y., y Wagener, T. (2017). Enhanced groundwater recharge rates and altered recharge sensitivity to climate variability through subsurface heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(11), 2842–2847. doi: 10.1073/pnas.1614941114

Haykin, S., York, N., San, B., London, F., Sydney, T., Singapore, T., ... Montreal, K. (2009). *Neural Networks and Learning Machines Third Edition* (3.ª ed.). Ontario: Pearson.

Hengl, T. (2018a). *Clay content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution.* OpenAIRE. doi: 10.5281/ZENODO.2525663

Hengl, T. (2018b). *Sand content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution.* OpenAIRE. doi: 10.5281/ZENODO.2525662

Hengl, T. (2018c). *Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m.* OpenAIRE. doi: 10.5281/ZENODO.2525817

Hengl, T., y Gupta, S. (2019). *Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution.* OpenAIRE. doi: 10.5281/ZENODO.2784001

Hennermann, K., y Guillory, A. (2020). *ERA5: Data documentation* . Descargado 2020-03-31, de https://confluence.ecmwf.int/display/CKB/ERA5{\%}3A+data+documentation{\#}ERA5:datadocumentation-HowtociteERA5

Hierro, L. (2016). *La revolución de los pozos "low cost".* Descargado 2019-05-25, de https://bit.ly/2Op2PeW

Horton, R. (1945). Erosional development of streams and their drainage basins: hydrophyisical approach to quantitative morphology. *Bulletin of the geological society of America*, *56*, 275–370. Descargado de https://bit.ly/2OuN4D8 doi: 10.1130/0016-7606(1945)56[275:EDOSAT]2.0.CO;2

Huayta, M. (2018). *Cavan pozo por casi S/5 millones y no encuentran agua.* Descargado 2019-05-24, de https://bit.ly/2nsxLQr

INEI. (2018). *Acceso a agua para consumo humano.* Lima. Descargado 2019-05-14, de https://bit.ly/2MlDPCD

INGEMMET. (2019). *Geología: estructuras.* Descargado 2019-09-19, de https://bit.ly/2KeoT77

Israil, M., Al-hadithi, M., y Singhal, D. C. (2006). Application of a resistivity survey and geographical information system (GIS) analysis for hydrogeological zoning of a piedmont area, Himalayan foothill region, India. *Hydrogeology Journal*, *14*(5), 753–759. doi: 10.1007/s10040-005-0483-0

Jacob, C. E. (1947). Drawdown Test to Determine Effective Radius of Artesian Well. *Transactions of American Society of Civil Engineers*, *1*(112), 1047–1064.

Jha, M. K., Chowdary, V. M., y Chowdhury, A. (2010). Groundwater assessment in Salboni Block, West Bengal (India) using remote sensing, geographical information system and multi-criteria decision analysis techniques. *Hydrogeology Journal*, *18*(7), 1713–1728. doi: 10.1007/s10040-010-0631-z

Jimenez, J. (2017). *Abastecimiento de agua subterránea con fines de uso agrícola para el fundo La Empedrada agroindustrial La Punta S.A.C - Huaura - Lima* (Tesis Doctoral no publicada). Universidad Nacional Agraria La Molina.

Kakish, K., y Katimbo, A. (2017). *Evaluation of groundwater recharge potential using GIS – Case study at the Salmon river watershed* (Inf. Téc.). Descargado de https://bit.ly/2Iuo3nM

Kaliraj, S., Chandrasekar, N., y Magesh, N. S. (2014). Identification of potential groundwater recharge zones in Vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique. *Arabian Journal of Geosciences*, *7*(4), 1385–1401. doi: 10.1007/s12517-013-0849-x

Kebede, S. (2013). *Groundwater in Ethiopia: Features, numbers and opportunities*. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-30391-3

Kleinbaum, D. G., y Klein, M. (2010). Introduction to Logistic Regression. En (pp. 1–39). Descargado de https://bit.ly/2nqdjzw doi: 10.1007/978-1-4419-1742-3_1

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. En *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 784 LNCS, pp. 171–182). Springer Verlag. Descargado de http://link.springer.com/10.1007/3-540-57868-4{_}57 doi: 10.1007/3-540-57868-4_57

Kotchoni, D. O., Vouillamoz, J. M., Lawson, F. M., Adjomayi, P., Boukari, M., y Taylor, R. G. (2019). Relationships between rainfall and groundwater recharge in seasonally humid Benin: a comparative analysis of long-term hydrographs in sedimentary and crystalline aquifers. *Hydrogeology Journal*, *27*(2), 447–457. doi: 10.1007/s10040-018-1806-2

Lee, S., Hong, S. M., y Jung, H. S. (2018). GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. *Geocarto International*, *33*(8), 847–861. Descargado de https://www.tandfonline.com/doi/full/10.1080/10106049.2017.1303091 doi: 10.1080/10106049.2017.1303091

Lee, S., Song, K. Y., Kim, Y., y Park, I. (2012). Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. *Hydrogeology Journal*, *20*(8), 1511–1527. doi: 10.1007/s10040-012-0894-7

Lehner, B., Verdin, K., y Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos*, *89*(10), 93–94. doi: 10.1029/2008EO100001

Liniger, H., y Weintgartner, R. (2008). *Mountains and freshwater supply.* Descargado 2019-08-30, de https://bit.ly/2OtCsof

Liška, A., Kruszewski, G., y Baroni, M. (2018). Memorize or generalize? Searching for a compositional RNN in a haystack. En *Aegap.* Estocolmo: Universidad de Cornell. Descargado de http://arxiv.org/abs/1802.06467

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice* (Tesis Doctoral, Universidad de Cornell). Descargado de https://bit.ly/2oX2Z2z

Machiwal, D., Jha, M. K., y Mal, B. C. (2011). Assessment of Groundwater Potential in a Semi-Arid Region of India Using Remote Sensing, GIS and MCDM Techniques. *Water Resources Management*, *25*(5), 1359–1386. doi: 10.1007/s11269-010-9749-y

McDonald, R. I., Weber, K., Padowski, J., Flörke, M., Schneider, C., Green, P. A., ... Montgomery, M. (2014). Water on an urban planet: Urbanization and the reach of urban water infrastructure. *Global Environmental Change*, *27*(1), 96–105. doi: 10.1016/j.gloenvcha.2014.04.022

McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, *17*(7), 1425–1432. doi: 10.1080/01431169608948714

McLachlan, A., y Brown, A. (2006). *The ecology of sandy shores, secend edition.* London: Elsevier Ltd. Descargado de https://bit.ly/2VssEfx

Mendoza, Z., Santayana, T., y Grover, U. (2010). *Recursos hídricos subterráneos en Perú.* Descargado 2019-08-30, de https://bit.ly/2oX3auL

Minagri. (2018). *Autoridad Nacional del Agua inicia la perforación de 30 piezómetros en el acuífero Caplina en Tacna.* Descargado 2019-05-06, de https://bit.ly/2pPUxlT

Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., ... Tien Bui, D. (2020). The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena*, *187*, 104421. doi: 10.1016/j.catena.2019.104421

Muñoz, I. (2015). Adaptación y debilidad del Estado: el caso de la escasez de agua subterránea en Ica. *Revista De Ciencia Política Y Gobierno*, *2*(4), 47–68. Descargado de http://

revistas.pucp.edu.pe/index.php/cienciapolitica/article/view/14899

Naghibi, S. A., Ahmadi, K., y Daneshi, A. (2017). Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. *Water Resources Management*, *31*(9), 2761–2775. doi: 10.1007/s11269-017-1660-3

Naghibi, S. A., Pourghasemi, H. R., y Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, *188*(1), 1–27. doi: 10.1007/s10661-015-5049-6

Nampak, H., Pradhan, B., y Manap, M. A. (2014). Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *Journal of Hydrology*, *513*, 283–300. doi: 10.1016/j.jhydrol.2014.02.053

Nassif, S. H., y Wilson, E. M. (1975). The influence of slope and rain intensity on runoff and infiltration. *Hydrological Sciences Bulletin*, *20*(4), 539–553. doi: 10.1080/02626667509491586

Oh, H. J., Kim, Y. S., Choi, J. K., Park, E., y Lee, S. (2011). GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. *Journal of Hydrology*, *399*(3-4), 158–172. doi: 10.1016/j.jhydrol.2010.12.027

Olden, J. D., Lawler, J. J., y Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *Quarterly Review of Biology*, *83*(2), 171–193. doi: 10.1086/587826

Ozdemir, A. (2011). Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *Journal of Hydrology*, *405*(1-2), 123–136. doi: 10.1016/j.jhydrol.2011.05.015

Parker, S. P. (1984). *McGraw-Hill concise encyclopedia of science & technology*. McGraw-Hill.

Pino, E., y Coarita, F. (2018). Caracterización hidrogeológica para determinar el deterioro de la calidad del agua en el acuifero la yarada media. *Revista de Investigaciones Altoandinas -Journal of High Andean Research*, *20*(4), 477–490. doi: 10.18271/ria.2018.424

Pourtaghi, Z. S., y Pourghasemi, H. R. (2014). GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeology Journal*, *22*(3), 643–662. doi: 10.1007/s10040-013-1089-6

QGIS Development Team. (2020). *QGIS Geographic Information System.* Open Source Geospatial Foundation. Descargado de http://qgis.org

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. doi: 10.1007/bf00116251

Rahmati, O., Pourghasemi, H. R., y Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *Catena*, *137*, 360–372. doi: 10.1016/j.catena.2015

.10.010

Rathay, S. Y., Allen, D. M., y Kirste, D. (2018). Response of a fractured bedrock aquifer to recharge from heavy rainfall events. *Journal of Hydrology*, *561*, 1048–1062. doi: 10.1016/j.jhydrol.2017.07.042

Razandi, Y., Pourghasemi, H. R., Neisani, N. S., y Rahmati, O. (2015). Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Science Informatics*, *8*(4), 867–883. doi: 10.1007/s12145-015-0220-8

Rebelo de Sá, C. (2019). Variance-Based Feature Importance in Neural Networks. En *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 11828 LNAI, pp. 306–315). Springer. doi: 10.1007/978-3-030-33778-0_24

Robinson, N. P., Allred, B. W., Jones, M. O., Moreno, A., Kimball, J. S., Naugle, D. E., ... Richardson, A. D. (2017). A dynamic landsat derived normalized difference vegetation index (NDVI) product for the conterminous United States. *Remote Sensing*, *9*(8). doi: 10.3390/rs9080863

Rohde, M. (2014). *Recharge: Groundwater's Second Act.* Descargado 2019-08-30, de `https://stanford.io/2LUSl5n`

Running, S., Mu, Q., y Zhao, M. (2017). *MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006.* doi: https://doi.org/10.5067/MODIS/MOD16A2.006

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., y Schisterman, E. F. (2008). Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*, *50*(3), 419–430. doi: 10.1002/bimj.200710415

Salazar, E. (2018). *Los dueños del agua en el desierto: Agroexportadoras y azucareras.* Descargado 2019-05-24, de `https://bit.ly/2GPRsdj`

Santillán, T. (2018). *Pozos tubulares: ¿solución para Cajamarca?* Descargado 2019-05-25, de `https://bit.ly/35etRLZ`

Schober, P., Boer, C., y Schwarte, L. A. (2018). Correlation Coefficients. *Anesthesia & Analgesia*, *126*(5), 1763–1768. Descargado de `http://journals.lww.com/00000539-201805000-00050` doi: 10.1213/ANE.0000000000002864

Sedapal. (2017). *Gestión hídrica.* Descargado 2019-05-24, de `https://bit.ly/2VssRiP`

Shenga, Z. D., Baroková, D., y Šoltész, A. (2018). Numerical modeling of groundwater to assess the impact of proposed railway construction on groundwater regime. *Pollack Periodica*, *13*(3), 187–196. doi: 10.1556/606.2018.13.3.18

Spiegel, M. (2017). *A Machine Learning Approach to Predicting Groundwater Potential in Zambia from Geologic and Remotely Sensed Variables* (Tesis Doctoral no publicada). Princeton University.

Theobald, D. M., Harrison-Atlas, D., Monahan, W. B., y Albano, C. M. (2015). Ecologically-

Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning. *PLOS ONE*, *10*(12). Descargado de https://dx.plos.org/10.1371/journal.pone.0143619 doi: 10.1371/journal.pone.0143619

Thiem, G. (1906). *Hydrologische methoden (Métodos hidrológicos)* (Tesis de doctorado). Universidad de Stuttgart.

Todd, D. K., y Mays, L. W. (2005). *Groundwater hydrology*. Wiley.

Torres, C. (2006). *Procedimiento para la Prueba de Bombeo: Escalonado y de Larga Duración.* Descargado 2019-05-19, de https://bit.ly/2MpTHUv

Tovar, J. (2005). *El Agua Subterránea en el medio ambiente minero Peruano.* Descargado 2019-05-25, de https://bit.ly/2VoRBbz

Tügel, F., Houben, G. J., y Graf, T. (2016). How appropriate is the Thiem equation for describing groundwater flow to actual wells? *Hydrogeology Journal*, *24*(8), 2093–2101. Descargado de http://link.springer.com/10.1007/s10040-016-1457-0 doi: 10.1007/s10040-016-1457-0

USGS. (2019). *Fault.* Descargado 2019-06-24, de https://on.doi.gov/2LU6jou

Van Tonder, G. J., Botha, J. F., y Van Bosch, J. (2001). A generalised solution for step-drawdown tests including flow dimension and elasticity. *Water SA*, *27*(3), 345–354. doi: 10.4314/wsa.v27i3.4978

Walsh Perú. (2009). *Estudio de Impacto Ambiental y Social - Proyecto Nitratos del Perú* (Inf. Téc.). Lima: Minem.

Yang, K., y Trewn, J. (2004). *Multivariate statistical methods in quality management*. McGraw-Hill.

Anexo

Tabla de estratos geológicos del Perú

Código original	Abreviación	Nombre de formación geológica	Valor adoptado en el modelo
8	Pe-m	Paleógeno eocena - marino	1
16	Pp-c	Paleógeno paleocena - continental	2
40	Ki-mc	Cretáceo inferior - marino, continental	3
84	JsKi-vs	Jurásico superior, Cretáceo inferior - volcanosedimentario	4
127	Qh-c	Cuaternario holocena - continental	5
163	Nm-vs	Neógeno miocena - volcanosedimentario	6
238	Ji-m	Jurásico inferior - marino	7
265	Nm-vs	Neógeno miocena - volcanosedimentario	8
324	PN-c	Paleógeno, Neógeno - continental	9
419	D-m	Devoniano - marino	10
433	Ks-tn,gd	Cretáceo superior - plutones	11
453	Qp-vs	Cuaternario pleistocena - volcanosedimentario	12
455	Np-m	Neógeno pliocena - marino	13
463	Js-vs	Jurásico superior - volcanosedimentario	14
481	Ki-tn,di	Cretáceo inferior - plutones	15
508	CpPE-m	Carbonífero pennsylvaniano, Permiano - marino	16
508	CpPE-m	Carbonífero penssylvaniano, Permiano - marino	17
509	N-gd,tn	Neógeno - plutones	18
544	Ki-m	Cretáceo inferior - marino	19
582	Po-m	Paleógeno oligocena - marino	20
613	Np-vs	Neógeno pliocena - volcanosedimentario	21
621	Kis-vs	Cretáceo inferior, superior - volcanosedimentario	22
672	NP-esq,gn	Neoproterozoica - esquisto, gneis	23
675	NP-gn	Neoproterozoica - gneis	24
705	Jim-sie	Jurásico inferior, media - plutones	25
718	Ki-mgr,gd	Cretáceo inferior - plutones	26
719	Ki-and,ri	Cretáceo inferior - pórfidos y subvolcánicos	27
831	Ks-and,ri	Cretáceo superior - pórfidos y subvolcánicos	28
1119	JsKi-mc	Jurásico superior, Cretáceo inferior - marino, continental	29
1183	Nm-m	Neógeno miocena - marino	30

1229	Jm-m	Jurásico media - marino	31
1306	Qp-c	Cuaternario pleistocena - continental	32
1319	PN-vs	Paleógeno, Neógeno - volcanosedimentario	33
1323	Cm-c	Carbonífero mississippiano - continental	34
1370	NP-gr	Neoproterozoica - plutones	35
1383	Ks-c	Cretáceo superior - continental	36
1391	Po-c	Paleógeno oligocena - continental	37
1391	Po-c	Paleógeno paleocena - continental	38
1463	PEl-c	Permiano lopingiana - continental	39
1499	KP-mgr,gd	Cretáceo, Paleógeno - plutones	40
1632	P-m	Paleógeno - marino	41
1768	Ji-vs	Jurásico inferior - volcanosedimentario	42
1779	KsP-c	Cretáceo superior, Paleógeno - continental	43
1840	Ks-di,tn,gd	Cretáceo superior - plutones	44
1856	KP-and,ri	Cretáceo, Paleógeno - pórfidos y subvolcánicos	45
1865	Js-and,ri	Jurásico superior - pórfidos y subvolcánicos	46
1876	D-gr	Devoniano - plutones	47
2149	N-fon	Neógeno - fonolita	48
2439	Qp-m	Cuaternario pleistocena - marino	49
2549	C-tn,gd	Carbonífero - plutones	50
2601	Ks-m	Cretáceo superior - marino	51
2604	Kis-m	Cretáceo inferior, superior - marino	52
2608	PET-tn,gd	Permiano, Triásico - plutones	53
2632	SD-ms	Siluriano, Devoniano - metasedimentario	54
2757	Ks-gd	Cretáceo superior - plutones	55
2796	N-dms	Neógeno - domos de sal	56
2849	N-gr,mz	Neógeno - plutones	57
2862	PN-tn,gd	Paleógeno, Neógeno - plutones	58
2889	PN-and,ri	Paleógeno, Neógeno - pórfidos y subvolcánicos	59
2915	N-gb,di	Neógeno - plutones	60
2942	Qp-v	Cuaternario pleistocena - volcánico	61
2944	Q-v	Cuaternario - volcánico	62
3003	Js-m	Jurásico superior - marino	63
3094	Np-v	Neógeno pliocena - volcánico	64
3116	NP-gn	Neoproterozoica - gneis	65
3202	Ks-gb,di	Cretáceo superior - plutones	66
3239	PET-and,ri	Pérmiano, Triásico - pórfidos y subvolcánicos	67
3322	C-mgr,gr	Carbonífero - plutones	68
3380	Jms-di,gd	Jurásico media, superior - plutones	69
3380	Jms-di,gd	Jurásico medio, superior - plutones	70
3387	Ki-c	Cretáceo inferior - continental	71
3416	P-c	Paleógeno - continental	72

3546	PN-mgr,gr	Paleógeno, Neógeno - plutones	73
3581	Nmp-c	Neógeno miocena, pliocena - continental	74
3605	KP-tn,gd,di	Cretáceo, Paleógeno - plutones	75
3614	P-vs	Paleógeno - volcanosedimentario	76
3669	O-gd	Ordoviciano - plutones	77
3706	OS-di,tn,gr	Ordoviciano, Siluriano - plutones	78
3746	PET-mgr,gr	Permiano, Triásico - plutones	79
3746	PET-mgr,gr	Pérmiano, Triásico - plutones	80
3773	Nmp-v	Neógeno miocena, pliocena - volcánico	81
3808	Ks-mgr,gd	Cretáceo superior - plutones	82
3831	Ki-v	Cretáceo inferior - volcánico	83
3933	KsP-vs	Cretáceo superior, Paleógeno - volcanosedimentario	84
4013	EO-ms	Cambriano, Ordoviciano - metasedimentario	85
4019	Jim-mgr,di	Jurásico inferior, media - plutones	86
4051	E-ms	Cambriano - metasedimentario	87
4089	TsJi-m	Triásico superior, Jurásico inferior - marino	88
4098	N-and,ri	Neógeno - pórfidos y subvolcánicos	89
4115	MNP-gn	Meso, Neoproterozoica - gneis	90
4133	Pp-m	Paleógeno paleocena - marino	91
4232	O-ms	Ordoviciano - metasedimentario	92
4234	Js-c	Jurásico superior - continental	93
4289	Nm-v	Neógeno miocena - volcánico	94
4379	Cm-v	Carbonífero mississippiano - volcánico	95
4388	NQ-c	Neógeno, Cuaternario - continental	96
4388	NQ-v	Neógeno, Cuaternario - volcánico	97
4394	C-and,ri	Carbonífero - pórfidos y subvolcánicos	98
4418	NQ-c	Neógeno, Cuaternario - continental	99
4679	KP-tn,gd	Cretáceo, Paleógeno - plutones	100
5078	Nm-c	Neógeno miocena - continental	101
5086	Jm-vs	Jurásico media - volcanosedimentario	102
5090	Np-c	Neógeno pliocena - continental	103
5297	NQ-vs	Neógeno, Cuaternario - volcanosedimentario	104
6162	TJi-tn,gd,mgr	Triásico, Jurásico inferior - plutones	105