Facultad de Ciencias Geológicas

Universidad Complutense de Madrid

MÁSTER UNIVERSITARIO EN GEOLOGÍA AMBIENTAL

Curso 2023-2024

Salinidad de las aguas subterráneas en Mali: Predicción espacial mediante herramientas de inteligencia artificial y estimación de personas en riesgo.

Groundwater salinity in Mali: Spatial prediction using machine learning tools and estimation of people at risk.

ARENE MALAXETXEBARRIA BENGOETXEA

TUTORES DEL TRABAJO: VÍCTOR GÓMEZ-ESCALONILLA CANALES Y PEDRO MARTÍNEZ SANTOS

Facultad de Ciencias Geológicas
Universidad Complutense de Madrid

MÁSTER UNIVERSITARIO EN GEOLOGÍA AMBIENTAL

Curso 2023-2024

Salinidad de las aguas subterráneas en Mali: Predicción espacial mediante herramientas de inteligencia artificial y estimación de personas en riesgo.

Groundwater salinity in Mali: Spatial prediction using machine learning tools and estimation of people at risk.

ARENE MALAXETXEBARRIA BENGOETXEA

TUTORES DEL TRABAJO: VÍCTOR GÓMEZ-ESCALONILLA CANALES Y PEDRO MARTÍNEZ SANTOS

Fdo.:

GOMEZ-ESCALONILLA CANALES VICTOR - 48283243X

Firmado digitalmente por GOMEZ-ESCALONILLA CANALES VICTOR - 48283243X
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-48283243X, givenName=VICTOR, sn=GOMEZ-ESCALONILLA CANALES, cn=GOMEZ-ESCALONILLA CANALES VICTOR - 48283243X
Fecha: 2024.07.29 13:50:43 +02'00'

MARTINEZ SANTOS PEDRO - DNI 50739588R

Firmado digitalmente por MARTINEZ SANTOS PEDRO - DNI 50739588R
Nombre de reconocimiento (DN): c=ES, o=UNIVERSIDAD COMPLUTENSE DE MADRID, ou=CERTIFICADO ELECTRONICO DE EMPLEADO PUBLICO, serialNumber=IDCES-50739588R, sn=MARTINEZ SANTOS, givenName=PEDRO, cn=MARTINEZ SANTOS PEDRO - DNI 50739588R
Fecha: 2024.07.29 16:29:42 +02'00'

Facultad de Ciencias Geológicas

Universidad Complutense de Madrid

DECLARACIÓN DE NO PLAGIO

ARENE MALAXETXEBARRIA BENGOETXEA con NIF 46369451W, estudiante de Máster en Geología Ambiental en la Facultad de Ciencias Geológicas de la Universidad Complutense de Madrid en el curso 2023-2024, como autora del trabajo de fin de máster titulado “Salinidad de las aguas subterráneas en Mali: Predicción espacial mediante herramientas de inteligencia artificial y estimación de personas en riesgo” y presentado para la obtención del título correspondiente, cuyos tutores son: VÍCTOR GÓMEZ-ESCALONILLA CANALES y PEDRO MARTÍNEZ SANTOS.

DECLARO QUE: El trabajo de fin de máster que presento está elaborado por mí y es original. No copio, ni utilizo ideas, formulaciones, citas integrales e ilustraciones de cualquier obra, artículo, memoria, o documento (en versión impresa o electrónica), sin mencionar de forma clara y estricta su origen, tanto en el cuerpo del texto como en la bibliografía. Así mismo declaro que los datos son veraces y que no he hecho uso de información no autorizada de cualquier fuente escrita de otra persona o de cualquier otra fuente. De igual manera, soy plenamente consciente de que el hecho de no respetar estos extremos es objeto de sanciones universitarias y/o de otro orden.

En Madrid, a 29 de julio de 2024.

Fdo.:

UNIVERSIDAD COMPLUTENSE MADRID

Declaración Responsable sobre Autoría y Uso Ético de Herramientas de Inteligencia Artificial (IA)

Yo, ARENE MALAXETXEBARRIA BENGOETXEA

Con DNI: 46369451 W

Declaro de manera responsable que el presente:

Trabajo de Fin de Máster (TFM)

Titulado “SALINIDAD DE LAS AGUAS SUBTERRÁNEAS EN MALI: PREDICCIÓN ESPACIAL MEDIANTE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL Y ESTIMACIÓN DE PERSONAS EN RIESGO”

es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo conlos principios éticos y las normas de integridad vigentes en la comunidad académica y, más específicamente, en la Universidad Complutense de Madrid.

Soy, pues, autor del material aquí incluido y, cuando no ha sido así y he tomado el material de otra fuente, lo he citado o bien he declarado su procedencia de forma clara -incluidas, en su caso, herramientas de inteligencia artificial-. Las ideas y aportaciones principales incluidas en este trabajo, y que acreditan la adquisición de competencias, son mías y no proceden de otras fuentes o han sido reescritas usando material de otras fuentes.

Asimismo, aseguro que los datos y recursos utilizados son legítimos, verificables y han sido obtenidos de fuentes confiables y autorizadas. Además, he tomado medidas para garantizar la confidencialidad y privacidad de los datos utilizados, evitando cualquier tipo de sesgo o discriminación injusta en el tratamiento de la información.

En Madrid a 29 DE JULIO DE 2024

Fdo.:

AGRADECIMIENTOS

A Víctor Gómez-Escalonilla, por toda la ayuda y paciencia. Muy contenta de que personas como tú vayan a formar parte de la nueva generación de investigadores y docentes.

A todos y a cada uno de mis compañeros de clase, por el compañerismo y por todo lo que me habéis aportado este año, mucho más allá del ámbito académico.

A mis amigas de toda la vida, por todo el apoyo y las visitas *express* a Madrid.

Y a mis compañeros de piso, por hacer mi año en Madrid todavía mejor.

ÍNDICE

RESUMEN

En regiones áridas como el Sahel, donde las sequías son recurrentes, las reservas de agua subterránea son esenciales para el ser humano. En estos contextos, la elevada salinidad del agua puede suponer un problema para la calidad del agua potable y, por tanto, para la salud humana. En este trabajo se lleva a cabo la predicción espacial mediante herramientas de inteligencia artificial de la conductividad eléctrica de las aguas subterráneas de gran parte de la República de Mali. El resultado final de este trabajo, en forma de cartografía, puede constituir una valiosa herramienta a la hora de mejorar el acceso al agua potable. Para ello, se ha digitalizado una base de datos de 21.196 pozos de agua distribuidos por todo el país y se han recopilado datos acerca de 18 variables explicativas relacionadas con la conductividad eléctrica de las aguas subterráneas que incluyen factores geológicos, climáticos y topográficos, entre otros. Se han evaluado cuatro umbrales diferentes de conductividad eléctrica para discernir entre puntos positivos y puntos negativos. El procedimiento ha incluido un preprocesamiento de las variables explicativas, una fase de entrenamiento de los algoritmos de clasificación y, finalmente, una etapa de validación en la que se ha evaluado la capacidad predictiva de los modelos. Los algoritmos que han presentado un mejor rendimiento han sido empleados para elaborar las cartografías predictivas. Los umbrales de conductividad eléctrica de 500 y 800 µS/cm fueron los que arrojaron mejores resultados. Los resultados muestran que las probabilidades más altas de que el agua presente una elevada salinidad se encuentran al noroeste y sureste de la zona de estudio. Finalmente, y mediante un mapa de densidad de población, se ha podido observar que los principales núcleos urbanos afectados por el consumo de agua subterránea de alta salinidad son Kayes, Nioro, Niono, Mopti, Douentza y Anefif.

1.INTRODUCCIÓN

El agua constituye un recurso natural vital para el ser humano. La Declaración Universal de los Derechos Humanos reconoció por primera vez en el año 1948, el acceso al agua como un derecho humano fundamental para el desarrollo de un estándar de vida digno. El acceso al agua potable es imprescindible para la hidratación, la higiene, el saneamiento y la seguridad alimentaria (United Nations, 2021, 2002). Sin embargo, el concepto "acceso", no solo hace referencia a la existencia del recurso como tal, sino también al estado cuantitativo y cualitativo adecuado para el consumo humano y a su disponibilidad y accesibilidad. Para ello, el Programa Conjunto de la OMS y UNICEF (WHO/UNICEF *Joint Monitoring Programme*) para el Monitoreo del Abastecimiento de Agua, del Saneamiento y de la Higiene define una serie de términos que permiten evaluar el nivel de acceso a este recurso. El concepto fundamental es el de "fuente mejorada de agua potable", que se define como aquella que tiene el potencial de ofrecer agua potable segura y libre de contaminación. Por otro lado, se define como un "servicio básico de agua potable" al acceso a una fuente de agua mejorada en el hogar o con un tiempo de recogida y acarreo menor a 30 minutos. En África, en el año 2020 solamente el 39 % de la población tuvo acceso a una "fuente mejorada de agua potable" (UNICEF/WHO, 2022). Además, existe una notable diferencia entre el acceso al agua potable en las zonas rurales y en las zonas urbanas. En las zonas urbanas, aproximadamente 2 de cada 5 personas no disponen de una "fuente mejorada de agua potable", mientras que en las zonas rurales 3 de cada 4 carecen del acceso a una fuente de este tipo (UNICEF/WHO, 2022). En el año 2020, en la República de Mali, zona de estudio de este trabajo, solamente el 72% de la población rural disponía de un "servicio básico de agua potable", en comparación con el 96% de la población urbana. A su vez, en muchos países africanos existen grandes desigualdades en el acceso a servicios básicos entre la población más rica y la más pobre. En la República de Mali, en el año 2018, solo el 40 % de la población más pobre hizo uso de al menos "servicios básicos de agua potable", mientras que el 95 % de las personas más ricas tuvieron acceso a ellos (UNICEF/WHO, 2021).

El agua subterránea es el principal recurso hídrico para alrededor de 2.500 millones de personas alrededor del mundo y, además, sustenta el bienestar de muchos ecosistemas en estado crítico (Gleeson et al., 2012). Sin embargo, en los últimos años, la sobreexplotación de estos recursos y la creciente demanda de agua han puesto en peligro su sostenibilidad (Wada et al., 2010; Motevalli et al., 2019). Existe otra problemática relacionada con el aumento de la salinidad que pueden sufrir las aguas subterráneas y que, en último término, puede repercutir gravemente en la calidad del agua potable y poner en peligro la salud humana (Gholami et al., 2017). Esta problemática es especialmente significativa en regiones áridas como el Sahel, donde las sequías son recurrentes y los recursos hídricos superficiales son cada vez más

escasos. En este contexto, la predicción espacial de potenciales puntos de explotación de agua subterránea o la predicción de parámetros hidroquímicos, relacionados con la salinidad o la contaminación, puede servir como una herramienta de gran interés para mejorar el acceso al agua potable. En las últimas décadas, este tipo de cartografías relacionadas con las aguas subterráneas, se han desarrollado mediante la utilización de técnicas de *Machine Learning* (Haggerty et al., 2023; Thanh et al., 2022). Los estudios que emplean diferentes modelos de aprendizaje automático para predecir la salinidad de las aguas subterráneas se han incrementado en los últimos años (Masciopinto et al., 2017; Bourke et al., 2017; Pauw et al., 2017; Levanon et al., 2017; Delsman et al., 2018; Gil-Márquez et al., 2017). Gran parte de los estudios basados en inteligencia artificial se han enfocado en la utilización de redes neuronales artificiales para elaborar este tipo de cartografías (Huang & Foo, 2022; Banerjee et al., 2011, Akramkhanov & Vlek, 2012; Alagha et al., 2017; Barzegar & Moghaddam, 2016). Sin embargo, la naturaleza de caja negra de algunas redes neuronales provoca que sea difícil cuantificar la contribución y la relación de las diferentes variables explicativas respecto a la salinidad de las aguas subterráneas (Sahour et al., 2020). Otros estudios han hecho uso de modelos pertenecientes a la familia de algoritmos "basados en árboles" como *Random Forest* (Akter et al., 2021; Mosavi et al., 2021) o *Gradient Boosting* (Sahour et al., 2020). Este tipo de algoritmos, dada su naturaleza, permiten analizar, en cierto modo, la importancia de las diferentes variables, permitiendo así comprender el funcionamiento interno de los modelos (Gómez-Escalonilla, 2024).

1.1.Objetivos

El objetivo principal de este trabajo es la elaboración de una cartografía predictiva de la conductividad eléctrica de las aguas subterráneas para la República de Mali, exceptuando la región de Tombouctou y gran parte de la región de Segou. Para ello, se va a emplear un enfoque de aprendizaje automático y, más concretamente, un enfoque de clasificación binaria. Para el correcto cumplimiento de dicho objetivo, será necesario desarrollar una serie de variables explicativas espacialmente distribuidas, así como aplicar una serie de técnicas y procedimientos estándar de aprendizaje automático.

Otro objetivo, derivado del objetivo principal, es evaluar la población en riesgo por abastecerse de aguas subterráneas que presentan una alta probabilidad de exceder un determinado umbral de conductividad eléctrica.

2.CARACTERIZACIÓN DE LA ZONA DE ESTUDIO

2.1. Contexto geográfico y climático

La República de Mali es un país ubicado en el oeste de África y abarca una extensión de 1.240.192 km^2. Limita al norte con Argelia, al oeste con Mauritania y Senegal, al sur con Guinea y Costa de Marfil y al este con Burkina Faso y Níger (Figura 1) (Oficina de Información Diplomática del Ministerio de Asuntos Exteriores, 2023).

La población total del país se encuentra cerca de los 24 millones de personas, equivalente al 0,29 % de la población total mundial. Tiene una densidad de población de 19 personas por km^2 y el 44 % de la población total vive en zonas urbanas (United Nations, 2022).

La República de Mali se compone de 8 regiones administrativas y un distrito capitalino (Bamako) (Figura 1). Estas unidades territoriales llevan el nombre de la ciudad principal de cada región. A su vez, cada región se encuentra dividida en círculos (*cercle*, en francés) resultado de la reagrupación de varias comunas. Una comuna rural es el resultado de la agrupación de una serie de aldeas. Esta información es importante, puesto que la base de datos de puntos de agua se encuentra estructurada por aldeas (Gómez-Escalonilla, 2024).

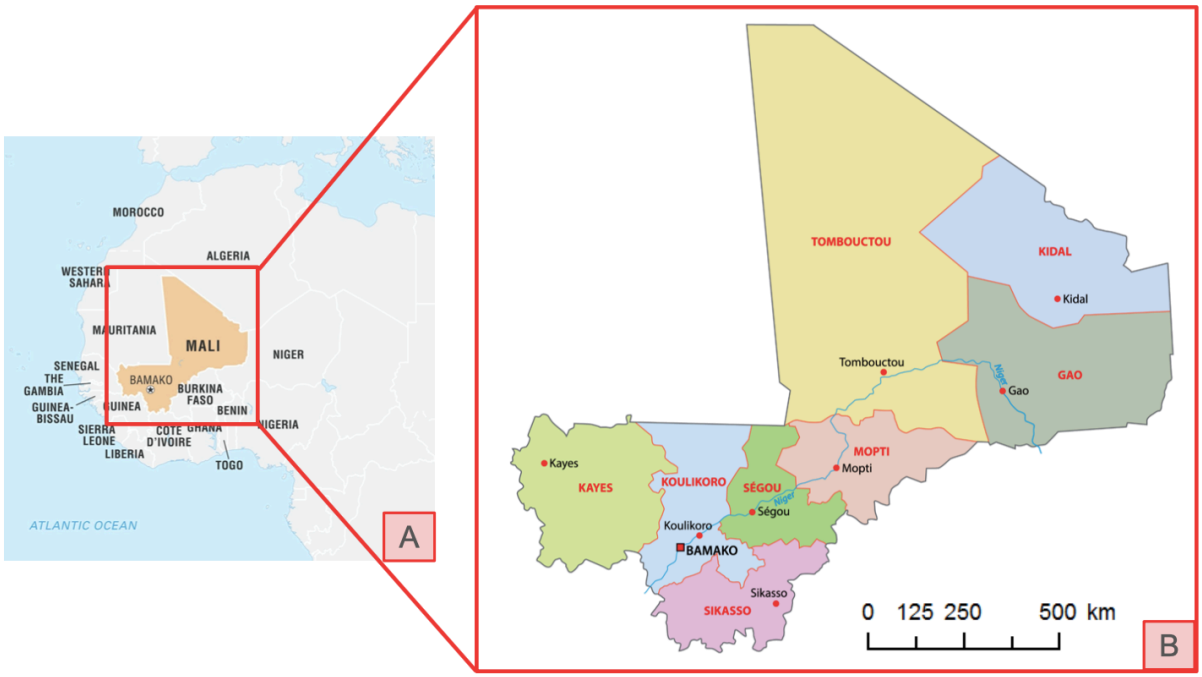

Figura 1: A) Mapa de la ubicación geográfica del país de Mali en África y sus respectivas fronteras terrestres. B) Mapa de las regiones administrativas de Mali. (Modificado de Imperato et al., 2024).

Desde un punto de vista topográfico, el paisaje está dominado por llanuras y mesetas (Díaz-Alcaide et al., 2017). El punto más bajo del país se encuentra en el río Senegal, a unos 20 m de altitud, mientras que el punto más alto del país es el pico Hombori Tondo con 1.153 m de altitud, situado en el centro de Mali (Figura 2). La frontera con Argelia al noreste también se considera zona de tierras altas (Traore et al., 2018).

Entre los cuerpos de agua superficial destacan los ríos Níger y Senegal, aunque también existen pequeños lagos como el lago Niangay o el lago Faguibine (Figura 2). El río Níger se considera el río más largo de África occidental y fluye hacia el noreste del país a través del delta interior (Smedley, 2022). El delta interior, considerado una zona llana con tributarios, pantanos y pequeños lagos, se convierte en un humedal de aproximadamente 30.000 km^2 durante la estación húmeda. Sin embargo, alrededor de dos tercios de la superficie total del país están clasificados como desérticos o semidesérticos (Smedley, 2022).

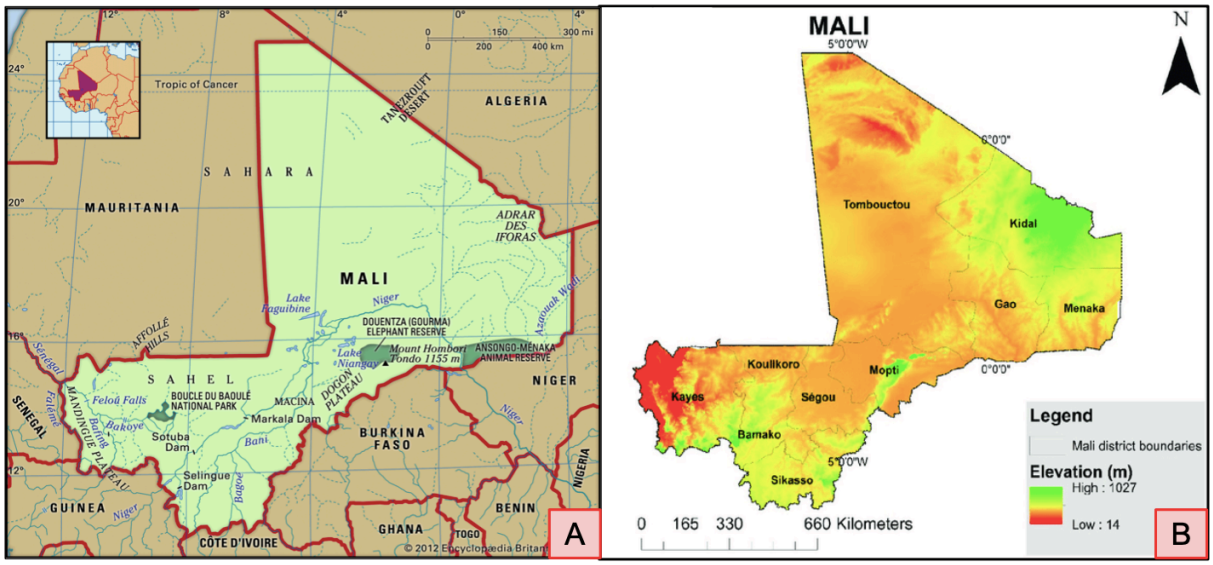

Figura 2: A) Mapa con los principales elementos geográficos y redes de drenaje superficiales de Mali (Modificado de Imperato et al., 2024). B) Mapa de elevación del terreno de Mali (Modificado de Attia et al., 2022).

Según la clasificación climática de Köppen-Geiger (Peel et al., 2007) el país se puede dividir en 3 zonas principales (Figura 3). El norte de Mali está dominado por el desierto del Sáhara, con un clima árido y cálido. Hacia el sur se transforma en la región semiárida del Sahel hasta llegar a la sabana tropical, el cual abarca el delta del río Níger (Traore et al., 2018).

El norte se caracteriza por una ausencia de lluvias y por una variación extrema de las temperaturas durante el día, las cuales ascienden a los 47ºC de día y descienden hasta los 4ºC de noche. La zona del Sahel se caracteriza por una media de 200-500 mm de precipitaciones anuales y las temperaturas medias se sitúan entre los 23 y 36ºC. En cambio, el sureste recibe una media de 500-1300 mm anuales y las temperaturas medias oscilan entre los 24 y 30ºC (Oficina de Información Diplomática del Ministerio de Asuntos Exteriores, 2023).

A pesar de las diferencias en términos de precipitación anual, los tres climas presentan una característica común. Las precipitaciones se concentran a lo largo de un periodo corto de tiempo, entre 3 y 6 meses (junio/julio-noviembre/diciembre) (Figura 4). El resto del año las precipitaciones son prácticamente inexistentes, esto condiciona el acceso a los recursos hídricos superficiales y aumenta la dependencia de los recursos hídricos subterráneos.

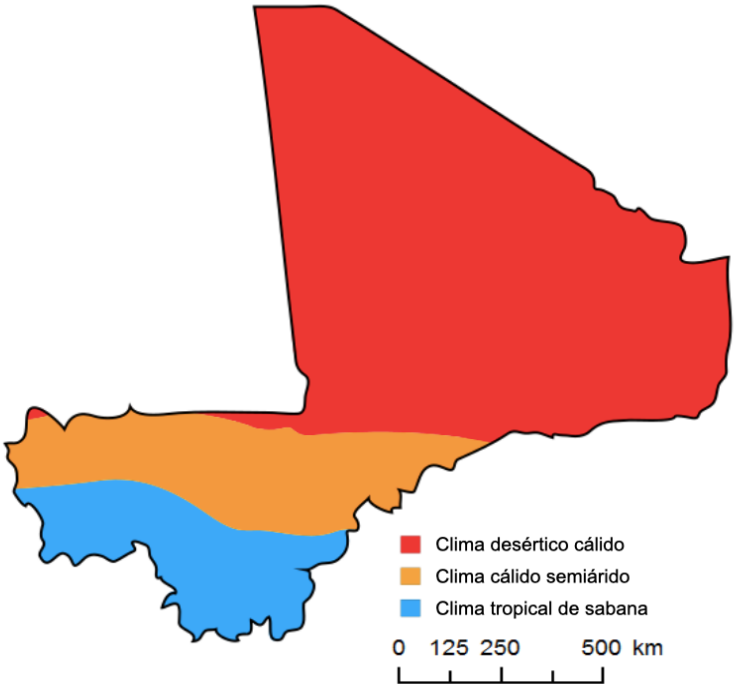

Figura 3: Mapa de zonas climáticas de Mali según la clasificación climática de Köppen Geiger (modificado de Jones & Harris, 2013).

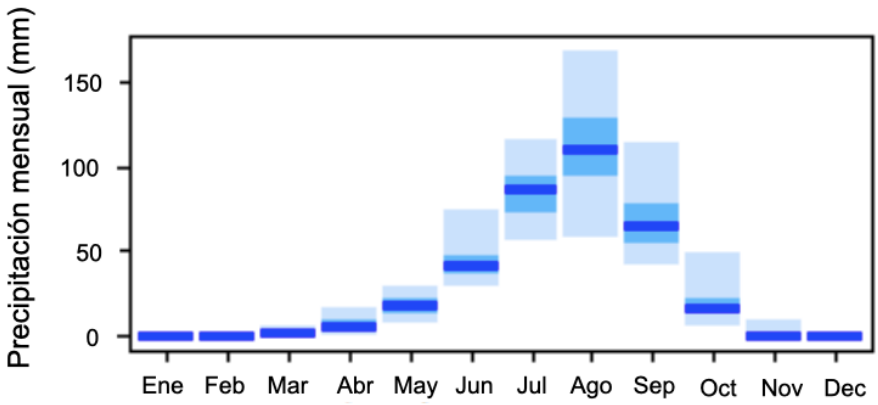

Figura 4: Las precipitaciones medias mensuales de Mali en mm. El color azul oscuro indica la precipitación media, el color azul medio los percentiles 25 y 75 y el color azul claro la precipitación mínima y máxima correspondiente a cada mes (modificado de Jones & Harris, 2013).

2.2. Contexto geológico e hidrogeológico

El basamento precámbrico de Mali está compuesto de dos núcleos cratónicos, las extensiones del cratón de África occidental y el escudo Tuareg, ambos formados durante la orogenia panafricana en el Neoproterozoico (Schlüter, 2006). El cratón de África occidental aflora en el oeste junto a la frontera senegalesa y en el sur, y el escudo Tuareg aflora en el este en las montañas de Adras des Iforas (Figura 5). El cratón de África occidental está compuesto principalmente por rocas metamórficas de origen volcánico-sedimentario. Por otro lado, aparecen rocas metasedimentarias, también de edad precámbrica, en forma de areniscas con un grado de metamorfismo medio-bajo, lutitas y calizas (Traore et al., 2018). Las rocas

volcánicas, que afloran en su mayoría al suroeste y al noroeste del país, incluyen basaltos y gabros de edad pérmica-triásica.

En la cuenca intracratónica de Taoudeni (Figura 5) y al este del país, afloran rocas sedimentarias de edades comprendidas entre el Paleozoico y el Cenozoico. Los materiales del Paleozoico son areniscas, calizas y pizarras. Por su parte, las rocas del Cretácico inferior, pertenecientes a la formación *Intercalaire* Continental, las conforman areniscas, conglomerados y arcillas. Finalmente, las rocas de edad cretácica superior-eocena se componen de secuencias de sedimentos marinos, en su mayoría con calizas en la base y areniscas y arcillas en la parte superior. En cambio, los materiales de edad miocena-pliocena son areniscas y arenas no consolidadas pertenecientes a la formación Continental Terminal. Por último, los sedimentos cuaternarios no consolidados, se localizan en las dunas de arena del desierto del Sahara y en los depósitos aluviales del río Níger (Traore et al., 2018).

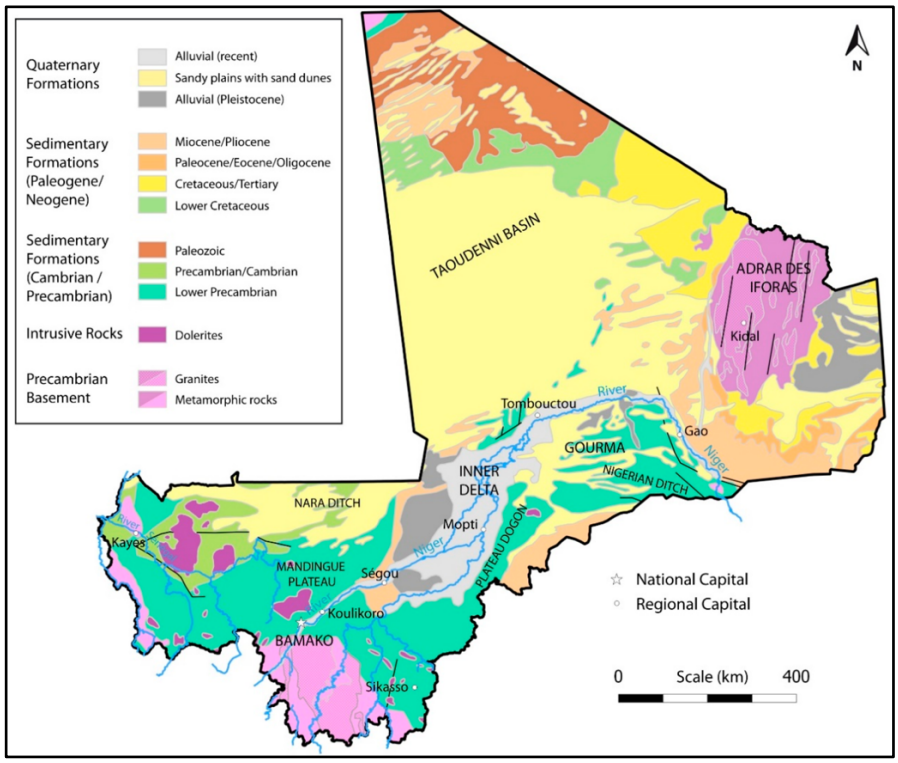

Figura 5: Mapa de las unidades geológicas principales de Mali (Díaz-Alcaide et al., 2017; modificado de Rasse, 2010).

La República de Mali, como se ha mencionado previamente, cuenta con escasos recursos hídricos superficiales, especialmente en el norte. Respecto a los recursos hídricos subterráneos, los principales acuíferos de Mali pueden distinguirse según el tipo de flujo dominante, aspecto altamente condicionado por el tipo de roca. Se diferencian aquellos en los que el flujo se produce principalmente a través de fracturas y aquellos con un flujo intergranular significativo. Los acuíferos fracturados son más frecuentes en el sur de Mali, en los materiales que conforman el basamento precámbrico. Los acuíferos intergranulares están asociados a sedimentos poco consolidados situados en las cuencas sedimentarias del este y el norte de

Mali. Además, sobre estos sistemas de acuíferos profundos, aparecen a menudo acuíferos superficiales del Cuaternario. Por su parte, las rocas ígneas intrusivas pueden actuar como una barrera o proporcionar vías preferentes para el flujo de agua subterránea (Traore et al., 2018).

En primer lugar, los acuíferos del basamento son generalmente semiconfinados y pueden dividirse en tres regiones: Los acuíferos del sur y suroeste se caracterizan por una precipitación elevada, una zona meteorizada de gran espesor y suelen estar drenados por el sistema del río Níger. En el oeste, región de Kayes, las precipitaciones son menores y el manto de alteración presenta un menor espesor. En la zona oriental, coincidiendo con la región saheliana, las precipitaciones son todavía más escasas y las aguas subterráneas se concentran a lo largo de las fracturas del basamento (Traore et al., 2018).

Los acuíferos sedimentarios consolidados, con flujo intergranular y a través de las fracturas, se sitúan en el sur y suroeste de Mali. Los metasedimentos precámbricos forman un acuífero multicapa, generalmente semiconfinado y de doble permeabilidad (Traore et al., 2018). Los acuíferos precámbricos, conformados por una alternancia de materiales de diferentes permeabilidades y situados en la región saheliana, se caracterizan por una recarga relativamente baja. Se trata, por tanto, de acuíferos discontinuos y las zonas más productivas están asociadas a fracturas en las capas de arenisca y caliza (Traore et al., 2018).

Entre los acuíferos no consolidados están los acuíferos continentales de edad cretácica inferior pertenecientes a la formación *Intercalaire* Continental descrita anteriormente. El acuífero del Cretácico Superior/Eoceno se encuentra en la franja occidental de la región del Adrar des Iforas y, en general, se considera un acuífero poco productivo. Por último, el acuífero compuesto por la formación Terminal Continental y los depósitos cuaternarios suprayacentes están en continuidad hidráulica y generalmente se consideran un único acuífero multicapa de alta productividad. El acuífero está formado por materiales asociados a las llanuras aluviales y, por ende, recibe una abundante recarga de agua superficial (Traore et al., 2018).

A nivel nacional, la República de Mali presenta un almacenamiento estimado de agua subterránea de alrededor de 27.100 km^3 y una productividad aproximada de entre 2 m^3/h y 20 m^3/h, dependiendo del tipo de acuífero (MacDonald et al., 2012). El acuífero de mayor productividad es el compuesto por la formación Continental Terminal y los depósitos cuaternarios suprayacentes, con una productividad estimada de entre 8-23 m^3/h aunque puede llegar a exceder los 100 m^3/h (Traore et al., 2018).

2.3. Estado cualitativo de las aguas subterráneas

Atendiendo a la calidad de las aguas subterráneas en la República de Mali, se puede establecer una división en tres regiones. El oeste y el sur del país se caracterizan por aguas subterráneas con bajos niveles de mineralización y, por tanto, con valores de conductividad eléctrica (CE) por debajo de 500 µS/cm. En la región central las aguas presentan una mineralización más elevada con valores de CE que oscilan entre los 300 y 1000 µS/cm. En cambio, en el norte y en el este de Mali, los acuíferos reciben menos recarga y las aguas subterráneas se encuentran más mineralizadas con valores de CE que superan los 1000 µS/cm y que incluso pueden alcanzar los 50000 µS/cm en algunas partes del desierto del Sahara (Traore et al., 2018).

En general, las aguas subterráneas no se ven afectadas notablemente por actividades antrópicas. Sin embargo, se han registrado algunos problemas de contaminación urbana y agrícola, sobre todo en acuíferos aluviales someros y en zonas meteorizadas (Traore et al., 2018).

3.MATERIALES Y MÉTODOS

3.1. Fundamentos teóricos

La conductividad eléctrica (CE) de un agua se define como la capacidad de conducir la corriente eléctrica y depende de la cantidad de iones disueltos en el agua. Se trata, también, de un indicativo del estado cualitativo del agua subterránea y, por tanto, de un parámetro a tener en cuenta al evaluar la calidad y potabilidad del recurso.

La conductividad eléctrica de las aguas subterráneas puede verse afectada por factores climáticos, factores geológicos e hidrogeológicos, variables edáficas o relacionadas con el tipo y los usos del suelo, factores topográficos y variables hidrogeomorfológicas (Golchin et al., 2016; Mosavi et al., 2020).

3.2. Base de datos de puntos de agua

La base de datos empleada en este trabajo fue proporcionada por la Dirección Nacional de Hidráulica de Mali (Direction Nationale de l'Hydraulique, 2010). La base de datos, digitalizada en gran medida como parte de este Trabajo de Fin de Máster, contiene información acerca de 21.196 pozos y sondeos distribuidos en 8.170 asentamientos en la República de Mali. La información disponible contiene el nombre del asentamiento o aldea donde están situados los sondeos y pozos, las demarcaciones administrativas correspondientes a cada aldea, el número de pozos positivos y negativos, la tasa de acierto, el caudal promedio en m^3/h (Tabla 1) y el número de pozos con un rango determinado de caudal (menor a 5 m^3/h, un caudal entre 5 y 10 m^3/h y un caudal mayor a 10 m^3/h). Un pozo se define como positivo cuando es capaz de captar una cantidad de agua subterránea suficiente como para satisfacer una necesidad específica, generalmente el consumo doméstico o el riego (Díaz-Alcaide et al., 2017). Generalmente, se utiliza el umbral de un caudal mayor a 0,5 m^3/h y se define como negativo cuando el pozo no supera dicho caudal (Foster et al., 2006). Atendiendo a este criterio, en la base de datos existen un total de 15.604 pozos positivos y 5.592 pozos negativos. La tasa de acierto se define como el porcentaje de perforaciones exitosas, refiriéndose a pozos y sondeos, con respecto al total de perforaciones.

Además, en algunos casos, también se incluye información acerca de la profundidad media del pozo (en metros), la profundidad media del nivel freático (en metros) y el valor medio de conductividad eléctrica (en µS/cm) del agua subterránea. La Tabla 1 muestra los valores medios, mínimos y máximos de la tasa de acierto, el caudal y la conductividad eléctrica de los pozos en las distintas regiones de Mali. La Figura 6 muestra la distribución espacial de las aldeas que contienen información sobre la conductividad eléctrica del agua subterránea. Estos

puntos serán utilizados para realizar las cartografías predictivas siguiendo el procedimiento mencionado en el apartado anterior.

Tabla 1: Valores medios, mínimos y máximos de la tasa de acierto, el caudal y la conductividad eléctrica de los pozos en las distintas regiones de Mali.

	Tasa de acierto (%)			Caudal (m^3/h)			Conductividad eléctrica (µs/cm)		
Región	**Promedio**	**Mínimo**	**Máximo**	**Promedio**	**Mínimo**	**Máximo**	**Promedio**	**Mínimo**	**Máximo**
Bamako	90,6	37,5	100	8,1	1	36,7	190,7	18	488
Gao	83,9	0	100	7,3	0,3	56	913,5	1	9999
Kayes	75,4	0	100	6,4	0,4	90	586,4	0	5137
Kidal	47,7	0	100	3,7	0,3	24	1011,3	166	2892
Koulikoro	79,8	0	100	4,6	0,5	60	339,3	16	10000
Mopti	86,7	0	100	7,9	0,3	100	599,4	7	5180
Segou	88,8	0	100	4,4	0,5	17	83,0	0	310
Sikasso	87,8	0	100	5,9	0	99	212,4	0	5880

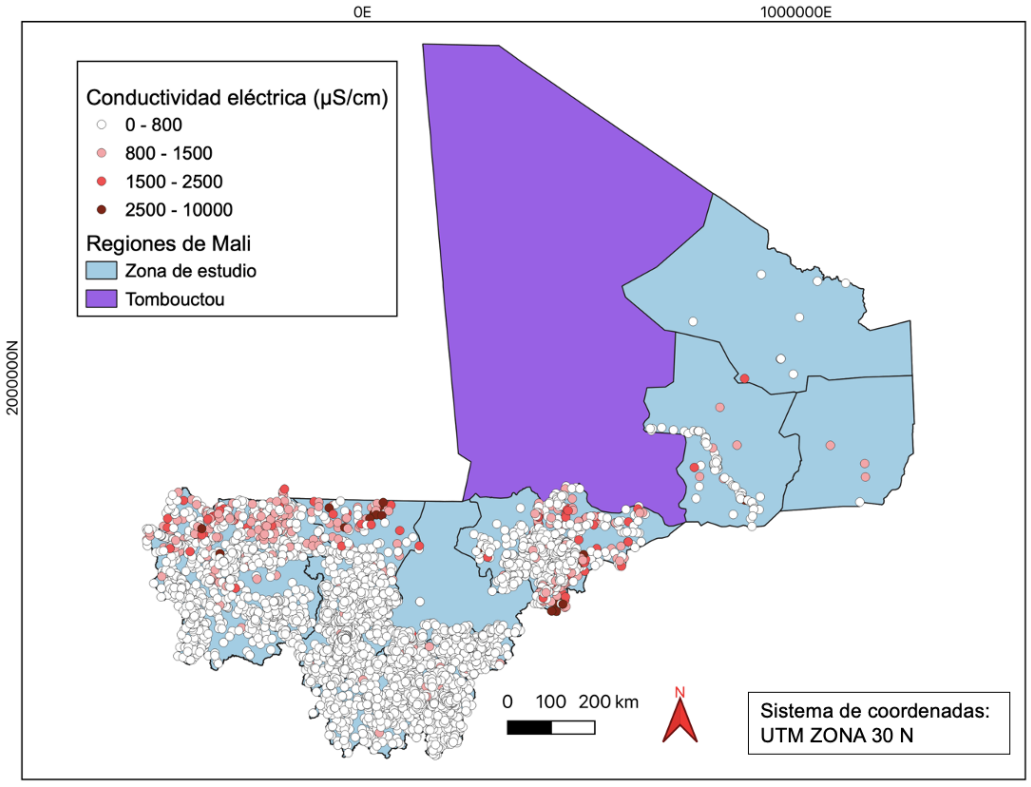

Figura 6: Distribución espacial de los datos de conductividad eléctrica medidos en el agua subterránea.

3.3. Variables explicativas

Entre los factores climáticos empleados se encuentran la temperatura, la evapotranspiración y la precipitación (Figura 7). Un incremento de la temperatura está asociado con una mayor tasa de evaporación, lo que puede provocar que el agua que se infiltre hacia los acuíferos tenga valores más elevados de CE (Araya et al., 2023). En relación con la precipitación, se considera que valores más elevados de esta variable pueden provocar una disminución de los

valores de CE del agua, ya que la precipitación es el principal condicionante de la recarga de los acuíferos, y puede disminuir la salinidad al diluir los iones disueltos en el agua subterránea (Geng & Boufadel, 2017). La información climatológica se ha obtenido de *Climate Engine* y, concretamente, del conjunto de datos climáticos *Terra Climate* (Abatzoglou et al., 2018; Huntington et al., 2017). La temperatura máxima promedio en grados centígrados, correspondiente al año 2010, se ha obtenido con una resolución espacial de 4 km. Además, se han obtenido los promedios de la precipitación total y la evapotranspiración real para el periodo que comprende desde el año 2000 hasta el año 2010. Ambas variables, referidas en mm/año, presentan una resolución de 4 km (Tabla 2).

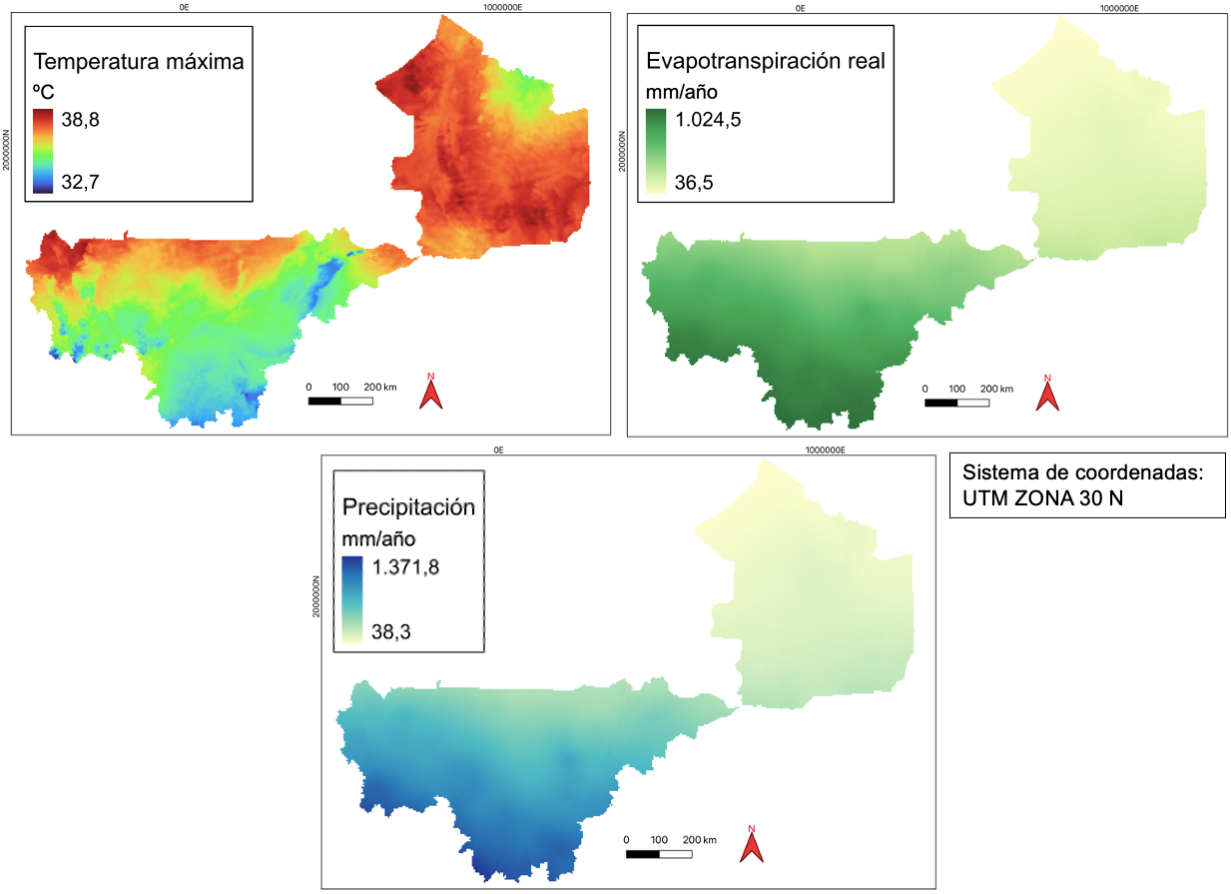

Figura 7: Variables explicativas usadas para predecir la conductividad eléctrica de las aguas subterráneas: Temperatura máxima, evapotranspiración real y precipitación.

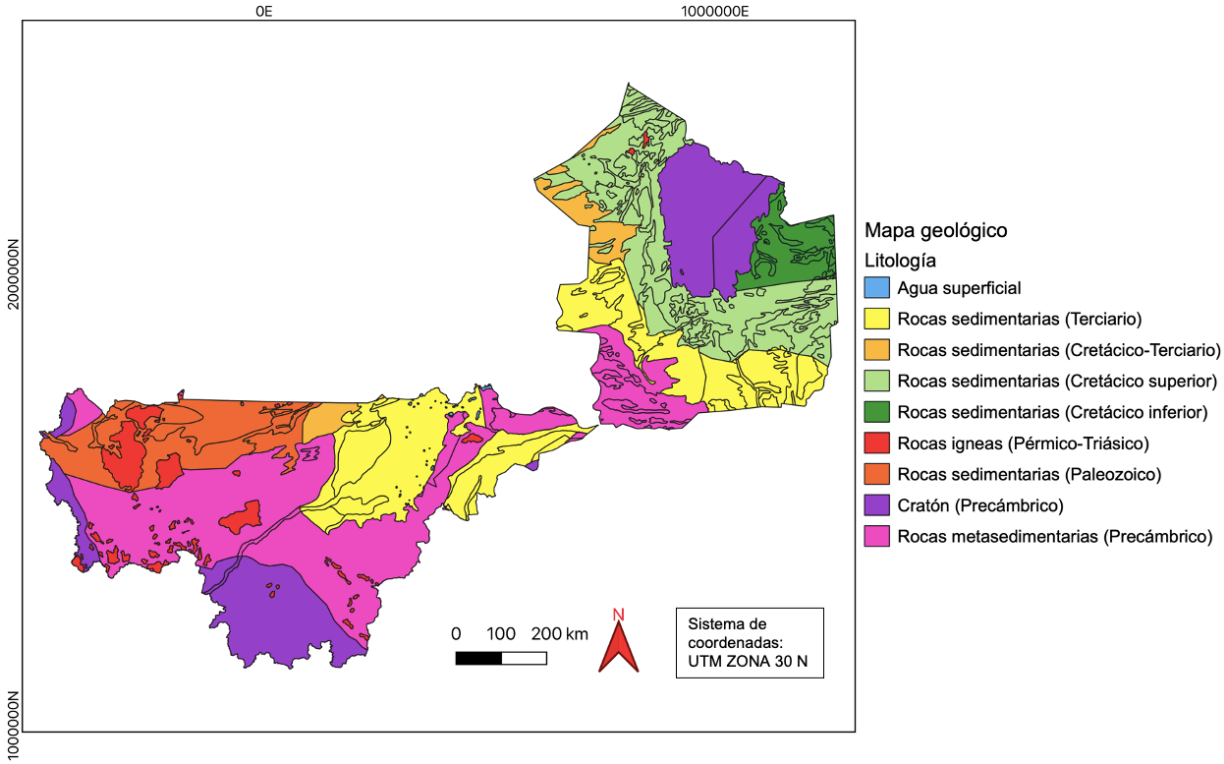

Figura 8: Variable explicativa usada para predecir la conductividad eléctrica de las aguas subterráneas: Geología (litología).

Las características geológicas e hidrogeológicas también tienen un papel fundamental en la CE de las aguas subterráneas. La litología es un factor importante (Figura 8), ya que la secuencia de rocas con las que está en contacto el agua en su transcurso por el subsuelo determinará las sales adquiridas por el agua. Además, la duración de contacto entre la roca y el agua estará relacionada con el grado de fisuración o permeabilidad del acuífero (Figura 9), factor que condiciona la velocidad de circulación del agua a través de los materiales. Por lo general, valores más elevados de permeabilidad pueden estar asociados a valores más bajos de CE en el agua subterránea. La información correspondiente a la geología e hidrogeología se ha adquirido de *Africa Groundwater Atlas*, disponible en *British Geological Survey* (BGS, 2024) como archivos *shapefile* con una escala de 1:5 millones. Dichos archivos contienen la litología y la edad de las principales unidades geológicas de la zona de estudio, así como el tipo de acuífero y su grado de productividad (Tabla 2).

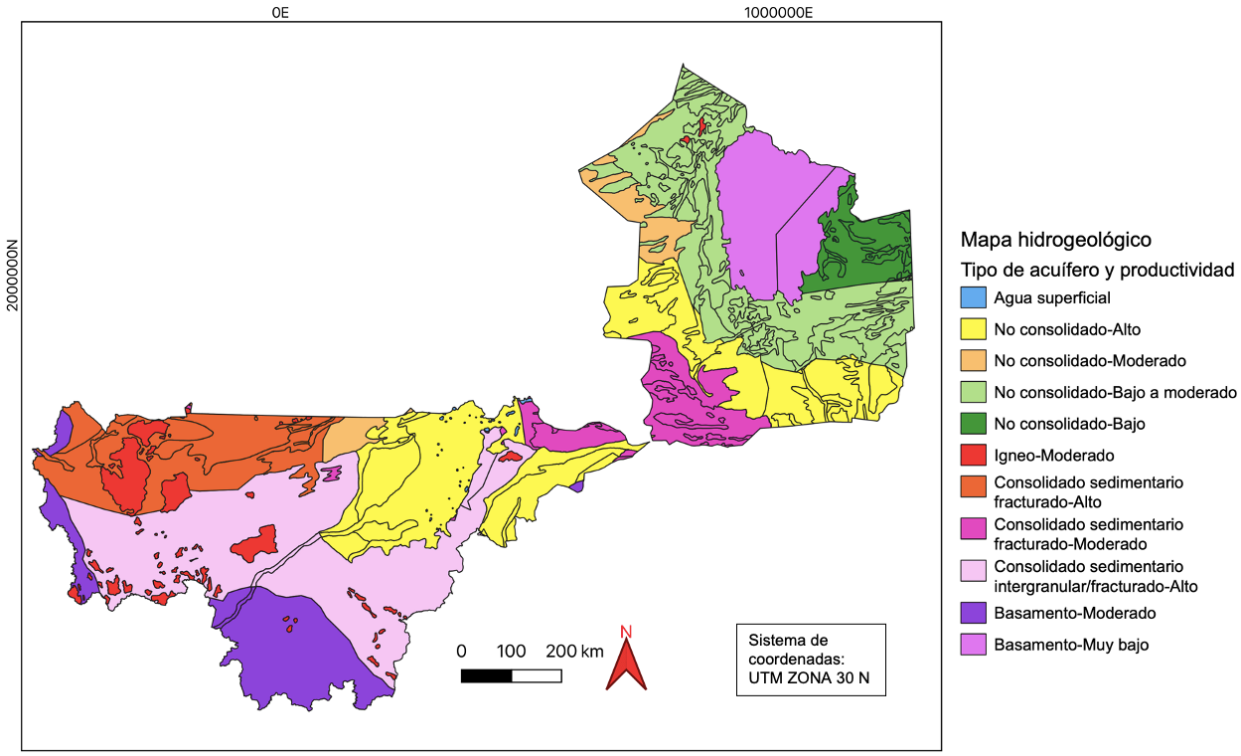

Figura 9: Variable explicativa usada para predecir la conductividad eléctrica de las aguas subterráneas: Hidrogeología (tipo de acuífero y su productividad).

El tipo de suelo (Figura 10) también puede influir en la acumulación de sales. Los suelos con texturas arcillosas a franco-arenosas muestran mayores niveles de salinidad que los suelos con texturas más gruesas debido a la baja permeabilidad y a la mayor capacidad de retención de fluidos que permite más tiempo para la evapotranspiración (Scanlon et al., 2010). Esta variable se ha obtenido a partir del mapa de tipo de suelo según la clasificación de los suelos de FAO del *Soil Atlas of Africa*, a escala de 1:3 millones, disponible en *European Soil Data Centre* (Dewitte et al., 2013). Además, se ha elaborado cartografías correspondientes al contenido del suelo en arena y en arcilla (g/kg) (Figura 10), con una resolución de 250 metros, a partir de los datos de *SoilGrids* (Poggio et al., 2021), con una resolución de 250 m. Para ello, se ha calculado el promedio del contenido en arena y arcilla entre dos intervalos de profundidad del suelo diferentes, entre 0-5 cm y 100-200 cm de profundidad (Tabla 2).

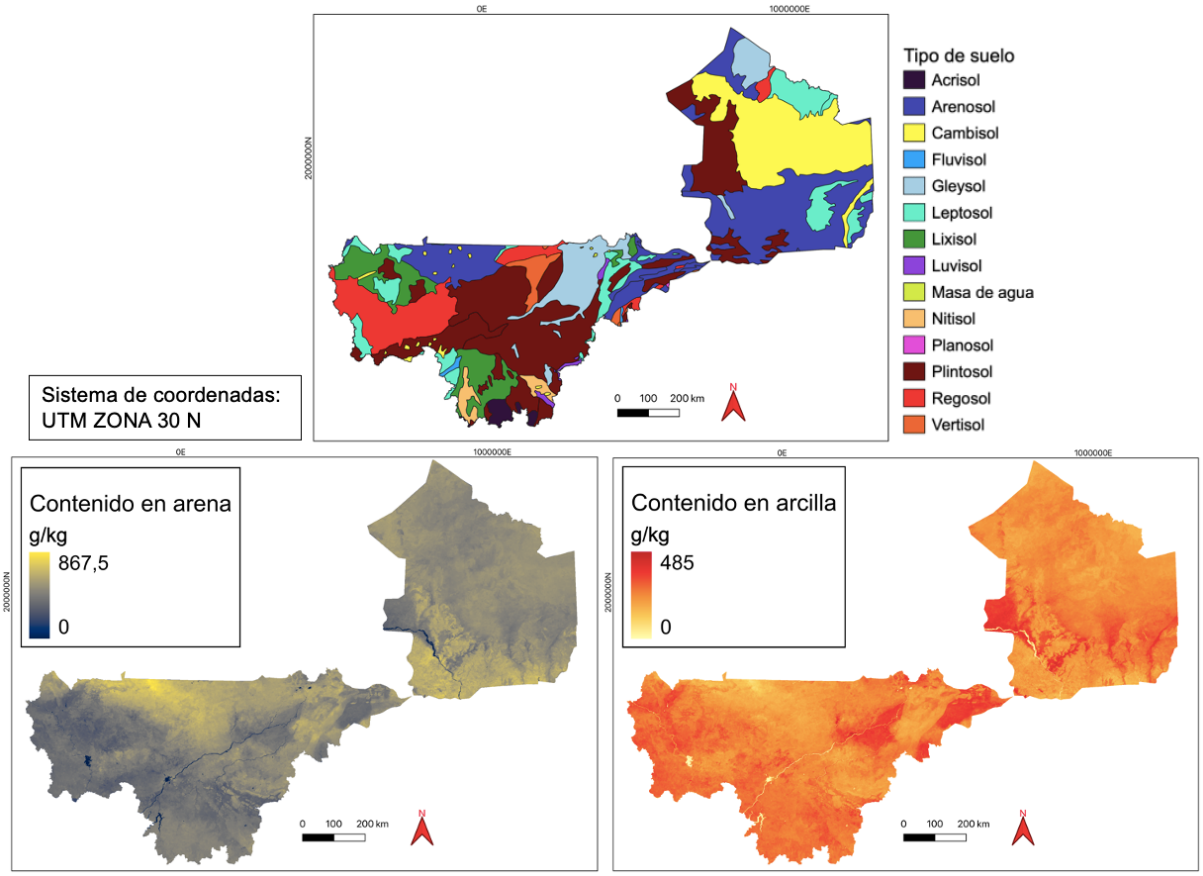

Figura 10: Variables explicativas usadas para predecir la conductividad eléctrica de las aguas subterráneas: Tipo de suelo, contenido en arena y contenido en arcilla.

Un factor, de origen antrópico, que puede influir en la CE de las aguas subterráneas son los usos del suelo. Aquellos relacionados con la agricultura pueden producir un aumento de la recarga de acuíferos a causa del riego, aumentando los valores de CE como efecto de la aplicación de fertilizantes. Otros usos del suelo relacionados con aprovechamientos antrópicos como los núcleos urbanos, con sus focos de contaminación asociados, pueden estar ligados a un aumento de la CE de las aguas subterráneas. La cobertera vegetal del suelo también tiene relación directa con la salinidad del agua subterránea, ya que la diferente capacidad de transpiración y profundidad de las raíces afecta enormemente a la escorrentía, evapotranspiración y drenaje subterráneo. Diferentes autores muestran que las plantaciones de árboles o bosques son capaces de acumular y captar más sal del subsuelo no saturado que los cultivos o plantas herbáceas (Nosetto et al., 2013; George et al., 1997). La cartografía de usos del suelo (Figura 11) de la zona de estudio se obtuvo de *European Space Agency-Climate Change Initiative* (ESA, 2010) con una resolución de 20 m y correspondiente al año 2016. Relacionados con factores superficiales, también se han obtenido dos índices, el Índice de Vegetación de Diferencia Normalizada (NDVI) y el Índice de Agua de Diferencia Normalizada (NDWI) de McFeeters (1996). Ambas cartografías (Figura 11) se han obtenido a partir de *Climate Engine* con una resolución de 240 metros para el final de la estación seca,

es decir, para el periodo de tiempo entre enero y mayo de 2010 (Tabla 2). El NDVI (Xie et al., 2008) es un indicativo de la densidad y el verdor de la cobertera vegetal y se obtiene calculando la diferencia de las intensidades de luz reflejada en el espectro del infrarrojo cercano y en el rango rojo del espectro (EOS Data Analytics, 2024). Mediante dicho cálculo se obtienen valores entre -1 y 1, donde los valores cercanos a 0 corresponden a arbustos y praderas y los valores cercanos a 1 representan en su mayoría bosques. En cambio, el NDWI indica el contenido de humedad de la vegetación o el suelo (Xu, 2006) y se calcula mediante la combinación de las bandas espectrales de verde visible e infrarrojo cercano. Los valores, al igual que el NDVI, varían entre -1 y 1 y los valores positivos indican la presencia de agua, siendo el valor 1 correspondiente a una superficie de agua (EOS Data Analytics, 2024).

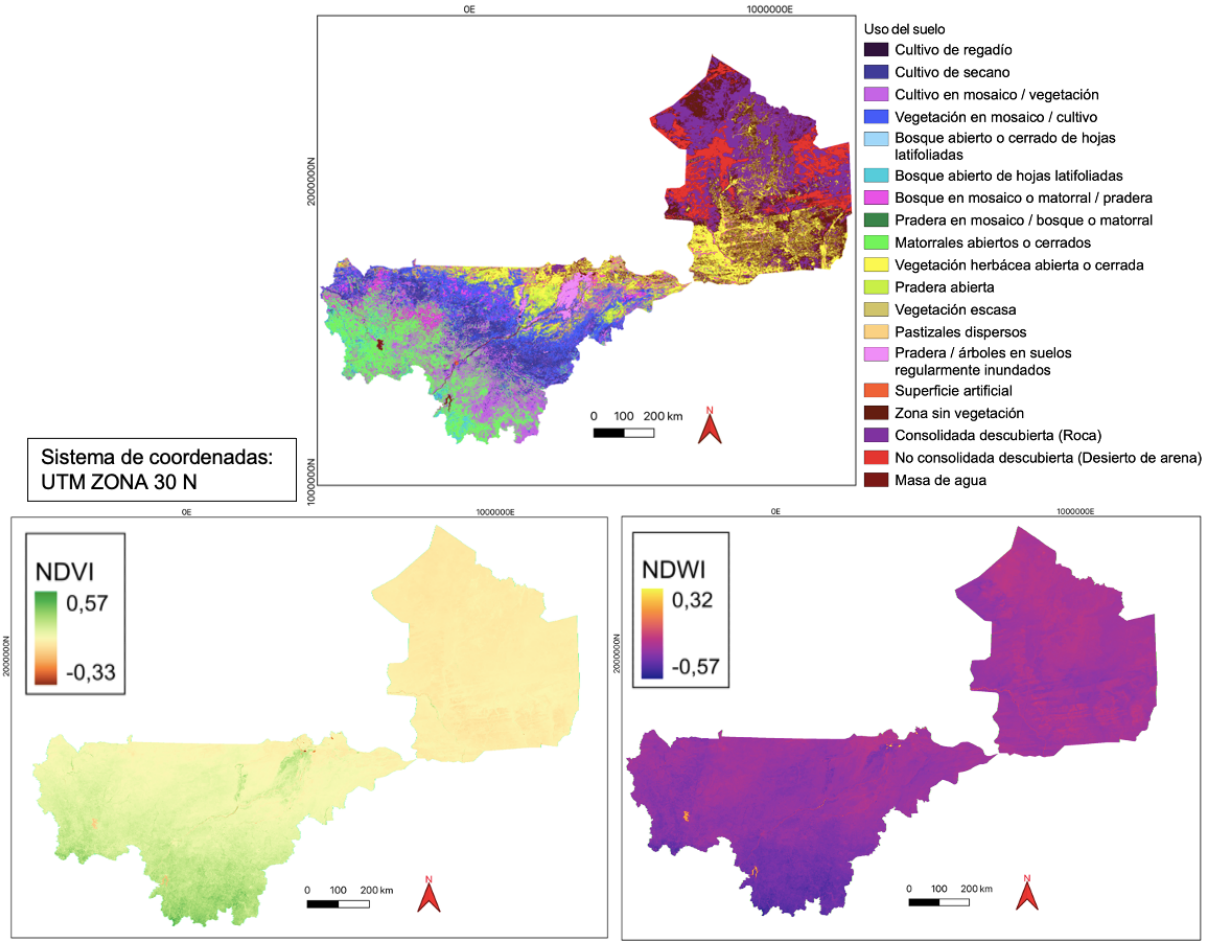

Figura 11: Variables explicativas usadas para predecir la conductividad eléctrica de las aguas subterráneas: Uso del suelo, NDVI y NDWI.

Los factores topográficos, como la elevación del terreno o la pendiente (Figura 12), también pueden influir en los valores de CE. En determinados contextos, la topografía condiciona la morfología y la profundidad del nivel freático (Salama et al., 1999). En zonas con un nivel freático somero, pueden aumentar los valores de evaporación y, por ende, los valores de CE (Geng & Boufadel, 2017). La pendiente, además, puede condicionar la recarga hacia los acuíferos (Gómez-Escalonilla, 2024). Generalmente, las zonas con mayor pendiente

favorecen la generación de escorrentía, mientras que las áreas con menores pendientes favorecen la infiltración y, en último término, están asociadas a una mayor recarga potencial (Farshae et al., 2014).

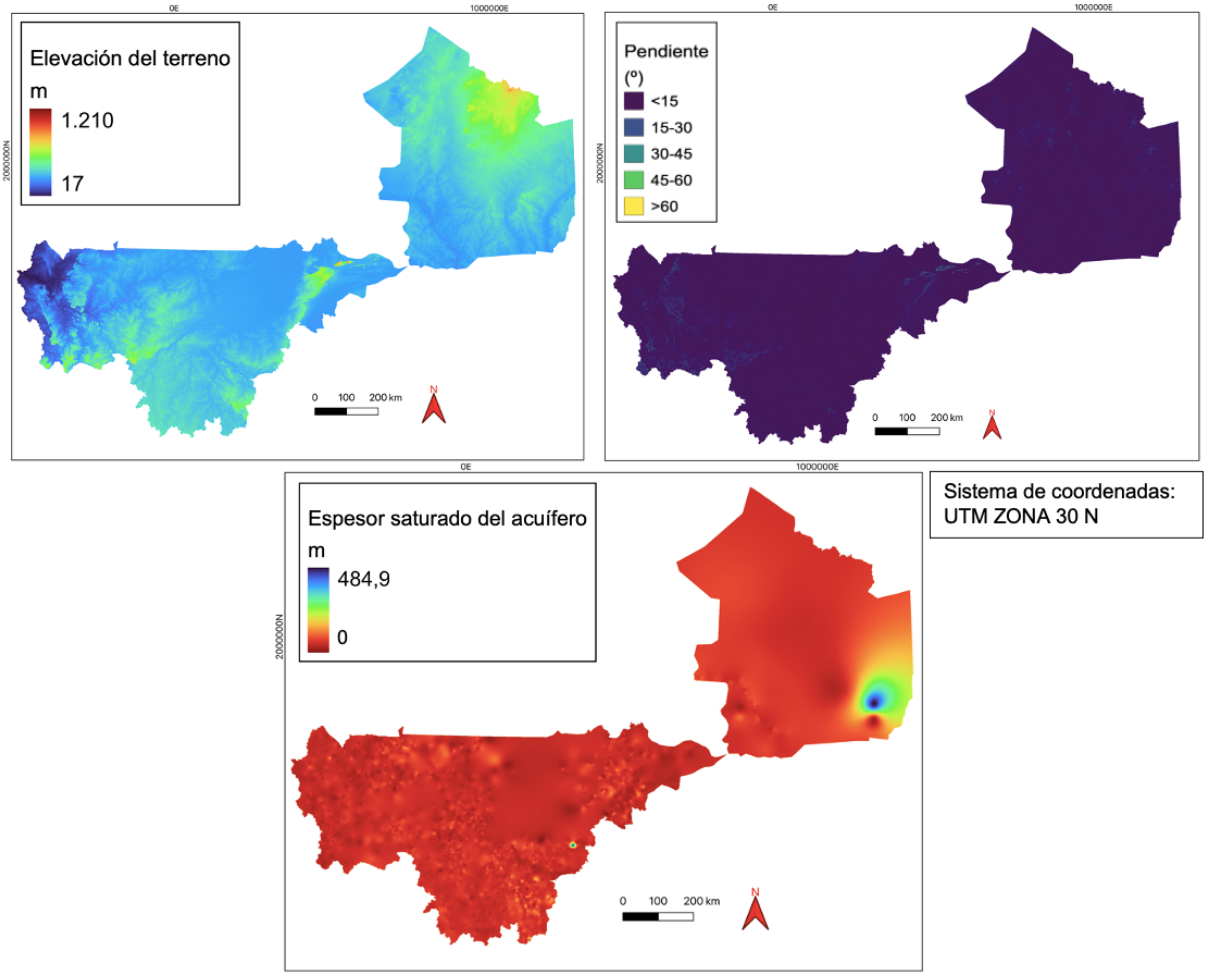

Figura 12: Variables explicativas usadas para predecir la conductividad eléctrica de las aguas subterráneas: Elevación del terreno, pendiente y espesor saturado del acuífero.

Estos factores se han obtenido a partir del modelo digital de elevación (DEM) en formato ráster (NASA Shuttle Radar Topography Mission, 2013) con una resolución de 1 arcseg (~30,53 metros). Posteriormente, a partir del DEM y mediante el empleo del software QGIS, se ha obtenido la pendiente del terreno en grados (Tabla 2). En zonas conformadas por materiales meteorizados sobre un basamento cristalino o sedimentario, como la de este caso de estudio, se puede asumir que, a grandes rasgos, la profundidad de perforación de los pozos se corresponde con la profundidad del basamento impermeable (Courtois et al., 2010; Gómez-Escalonilla, 2024). Por ello, mediante los datos proporcionados por la Dirección Nacional de Hidráulica de Mali (Direction Nationale de l'Hydraulique, 2010), se ha obtenido, en primer lugar, un mapa continuo de profundidad del basamento impermeable y profundidad del nivel freático mediante la herramienta *Multilevel B-spline interpolation*. Posteriormente, también

mediante esta herramienta de interpolación, y empleando en primer lugar la fórmula que se encuentra a continuación, se pudo obtener el espesor saturado del acuífero (Figura 12).

$$Espesor\ saturado\ del\ acuífero =$$

$$Profundidad\ del\ basamento\ impermeable - Profundidad\ del\ nivel\ freático$$

Finalmente, también se han considerado los factores hidrológicos, ya que las características de la red de drenaje pueden condicionar la salinidad de las aguas subterráneas. Los ríos perdedores recargan los acuíferos de agua dulce y, por tanto, la proximidad al río y el caudal transferido al acuífero se verá reflejado en los valores de CE de las aguas subterráneas, siendo estás más bajas cuanto más cerca esté el río y mayor sea el caudal transferido (Iris Rodríguez et al., 2010). Por otro lado, hay tener en cuenta la recarga de agua subterránea como un factor importante relacionado con la salinidad (Figura 13). La recarga está mayormente condicionada por la precipitación y por las características del suelo que a su vez controlan la capacidad de infiltración del agua (Gómez-Escalonilla et al., 2022). En regiones áridas a semiáridas, como el caso de la zona de estudio de este trabajo, el proceso dominante de recarga es concentrada. Esta se produce cuando el agua se infiltra desde fuentes de agua superficiales (ríos, lagos, ramblas, humedales) o zonas deprimidas (Cuthbert et al., 2019; MacDonald et al., 2021). Teniendo en cuenta lo anterior, se ha calculado la distancia a los canales fluviales (Figura 13) a partir del DEM empleando la herramienta de QGIS llamada *horizontal distance from channel network.* Además, se han obtenido los valores medios anuales de recarga de agua subterránea (Tabla 2) en mm/año para el periodo de tiempo entre 1970 y 2019 (McDonald et al., 2021; BGS, 2024).

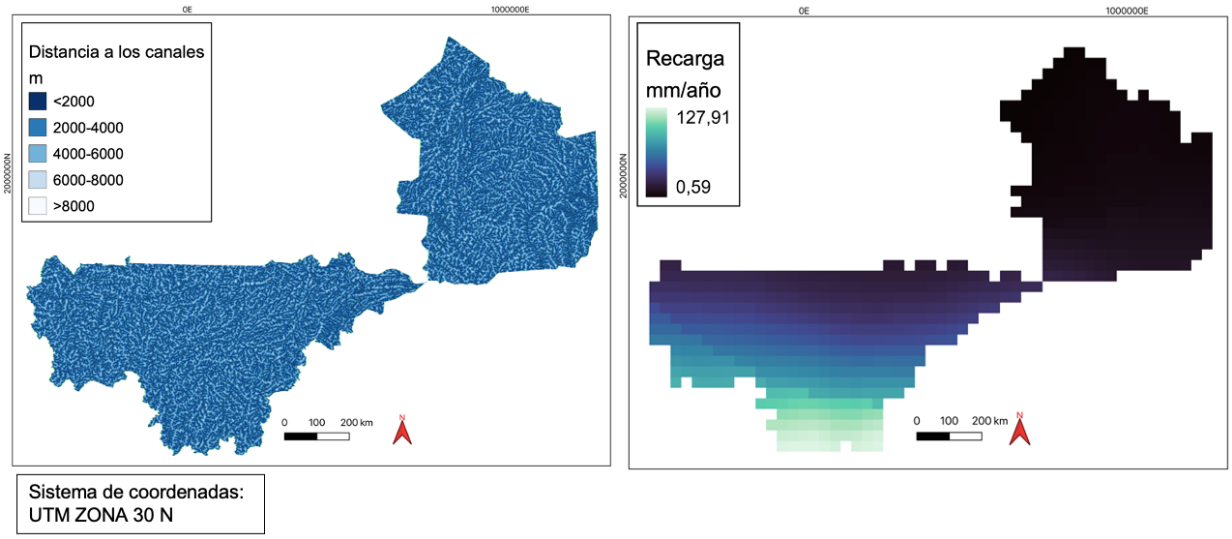

Figura 13: Variables explicativas usadas para predecir la conductividad eléctrica de las aguas subterráneas: distancia horizontal a los canales fluviales y recarga de agua subterránea.

Tabla 2: Las distintas variables explicativas y sus correspondientes unidades, resolución/escala, periodo de tiempo y base de datos/fuente.

	Variable	Unidad	Resolución/ Escala	Periodo de tiempo	Base de datos/Fuente
Factores climático	Temperatura máxima	ºC	4 km	1/01/2010-31/12/2010	Terra Climate - Climate Engine
	Precipitación total	mm/año	4 km	1/01/2000-31/12/2010	Terra Climate - Climate Engine
	Evapotranspiración real	mm/año	4 km	1/01/2000-31/12/2010	Terra Climate - Climate Engine
Factores superficiales	Tipo de suelo	-	1:3 M	-	Soil Atlas of Africa - European Soil Data Centre
	Uso del suelo	-	20 m	1/01/2016-31/12/2016	Sentinel-2A - ESA Climate Change Initiative
	Contenido en arena	g/kg	250 m	-	SoilGrids250m 2.0
	Contenido en arcilla	g/kg	250 m	-	SoilGrids250m 2.0
Factores topográficos	Elevación (DEM)	m	1 Arc-Second	-	SRTM - USGS (Earth Explorer)
	Pendiente	º	1 Arc-Second	-	Elaborado a partir del DEM
Factores hidrogeológicos e hidrológicos	Tipo de acuífero y productividad	-	1:5 M	-	BGS (Africa Groundwater Atlas)
	Distancia a los canales	m	1 Arc-Second	-	Elaborado a partir del DEM
	Profundidad del nivel freático	m	100 x 100 m	-	Direction Nationale de l'Hydraulique
	Profundidad del pozo	m	100 x 100 m	-	Direction Nationale de l'Hydraulique
	Espesor saturado del acuífero	m	100 x 100 m		Elaborado a partir de la base de datos de los pozos y del DEM
	Recarga del acuífero	mm/año	~ 32 km	1/01/1970-1/01/2020	BGS (Groundwater recharge in Africa from ground based measurements)
Factor geológico	Litología	-	1:5 M	-	BGS (Africa Groundwater Atlas)
Otros factores	NDVI	-	240 m	1/01/2010-31/05/2010	Landsat 5/7/8/9 - Climate Engine
	NDWI	-	240 m	1/01/2010-31/05/2010	Landsat 5/7/8/9 - Climate Engine

3.4. Enfoques de aprendizaje automático

El aprendizaje automático o *machine learning* es una rama de la inteligencia artificial que programa los ordenadores para que puedan aprender a partir de datos (Géron, 2019). Dependiendo de la base de datos y del objetivo del trabajo a realizar existen distintos tipos de enfoques de aprendizaje automático.

Por un lado, los algoritmos de aprendizaje supervisado que necesitan un conjunto de datos que incluyan tanto los valores de las variables explicativas como el valor de la variable objetivo. Los algoritmos, después de ser entrenados, van a conocer el comportamiento de las variables explicativas y esto les permitirá predecir la variable objetivo en los puntos donde no se conoce el estado de esa variable objetivo (Géron, 2019; Suthaharan, 2016). Por otro lado, los algoritmos de aprendizaje no supervisado trabajan con un conjunto de datos en el que no se conoce la variable objetivo, y el propósito es extraer información del conjunto de datos o agrupar las muestras atendiendo a distintas métricas de similitud (Gómez-Escalonilla, 2024). Dentro del aprendizaje supervisado, se encuentran algoritmos de clasificación y de regresión, según el tipo de variable objetivo. Los algoritmos de clasificación trabajan con variables objetivo con valores discretos y a su vez se pueden dividir en binarios, cuando solo hay dos clases, o multiclase, cuando hay más de dos. Por su parte, los algoritmos de regresión trabajan con valores continuos, en cuanto a la variable objetivo se refiere.

En este trabajo, los datos de entrada incluyen tanto valores de las variables explicativas como una variable objetivo. Por lo tanto, se ha empleado un enfoque de aprendizaje automático supervisado. En este caso, la variable objetivo es la conductividad eléctrica, la cual tiene valores continuos, sin embargo, se le han asignado valores de carácter binario en función de si supera o no ciertos umbrales de conductividad eléctrica, es decir, 1 si supera el umbral y 0 en caso de no superarlo, con el objetivo de facilitar las tareas de aprendizaje de los modelos. Los umbrales de conductividad eléctrica que se han evaluado son 500 µS/cm, 800 µS/cm, 1500 µS/cm y 2500 µS/cm. El umbral de 800 µS/cm corresponde al estándar de agua potable de buena calidad (WHO, 2011), el umbral de 1500 µS/cm corresponde al estándar de agua potable de sabor aceptable (WHO, 2011) y el umbral de 2500 µS/cm hace referencia a la salinidad máxima del agua que la población puede consumir cuando no existen otras opciones disponibles (Muthusi et al., 2007).

3.5. Procedimiento y software de clasificación supervisada

Para llevar a cabo las cartografías predictivas se ha empleado MLMapper v2.0, un código programado en Python que se utiliza como un plugin para la herramienta QGIS3 (Gómez-Escalonilla et al., 2022). Las tareas de aprendizaje automático incluyen diferentes fases (Figura 14) que se explican a continuación.

El primer paso consiste en generar una base de datos en un archivo de valores separados por comas (formato CSV) que contenga los puntos de pozos de agua subterránea, sus respectivas coordenadas, la variable objetivo asociada (en este caso la conductividad eléctrica con valores 0 y 1, según los umbrales previamente definidos) y las 18 variables explicativas asociadas con sus respectivos valores. A continuación, se divide la base de datos, por una parte, el 70 % de los datos que serán utilizados en las tareas de entrenamiento de los algoritmos y, por otra

parte, el 30 % restante que se utilizará para validar los modelos de *machine learning*. Posteriormente, comienza la fase de entrenamiento o aprendizaje de los algoritmos. Durante esta etapa, los modelos van a poder observar la variable objetivo y el valor de todas las variables explicativas (únicamente para el 70 % de la base de datos original) con el objetivo de encontrar patrones y asociaciones entre la conductividad eléctrica y las variables explicativas (Gómez-Escalonilla, 2024). Durante la etapa de entrenamiento, los hiperparámetros de los algoritmos son optimizados mediante una búsqueda aleatoria de optimización. En este trabajo se han realizado un total de 50 iteraciones durante esta fase de optimización. Después, durante la fase de validación, los algoritmos no son capaces de ver los valores de la variable objetivo y tiene que predecir dichos valores mediante los patrones aprendidos en la fase anterior. Posteriormente, se comparan los resultados arrojados por los algoritmos con los valores reales de la variable objetivo y se emplean una serie de métricas para conocer el grado de error y la fiabilidad de las predicciones. Finalmente, si las métricas arrojan resultados satisfactorios, los algoritmos pueden predecir la variable objetivo, relacionada con la conductividad eléctrica del agua subterránea, en una malla de puntos regulares distanciados unos de otros por 1000 m. Por un lado, se han realizado predicciones binarias, las cuales predicen el valor de la variable objetivo como 0 o 1 (superan o no superan un determinado umbral) y, por otro lado, se han realizado predicciones probabilísticas que, como su propio nombre indica, estiman la probabilidad de que el valor de la variable objetivo sea 0 o 1, es decir, que supere o no el umbral establecido.

Entre los modelos incluidos en el software MLMapper v2.0 se encuentran métodos de análisis discriminante lineal (LDA), la regresión logística (LRG), el clasificador k-vecinos (KNN), árboles de decisión (DTC), redes neuronales (MLP) y métodos *ensemble* basados en árboles como el *Random Forest Classifier* (RFC), *Gradient Boosting Classifier* (GBC), *Ada-Boost Classifier* y *Extra Trees Classifier* (ETC), entre otros (Gómez-Escalonilla et al., 2022).

Los algoritmos de aprendizaje automático son altamente complejos y, además, esto se combina con las complejas asociaciones que pueden existir entre las variables explicativas y la variable objetivo. Por lo tanto, desde un inicio no es posible conocer cuáles serán los algoritmos que se comportará mejor en una base de datos determinada. Por ello, se realizan una serie de ejecuciones con todos los algoritmos incluidos y, posteriormente, se seleccionan los algoritmos con mejor rendimiento y se descartan el resto (Gómez-Escalonilla, 2024).

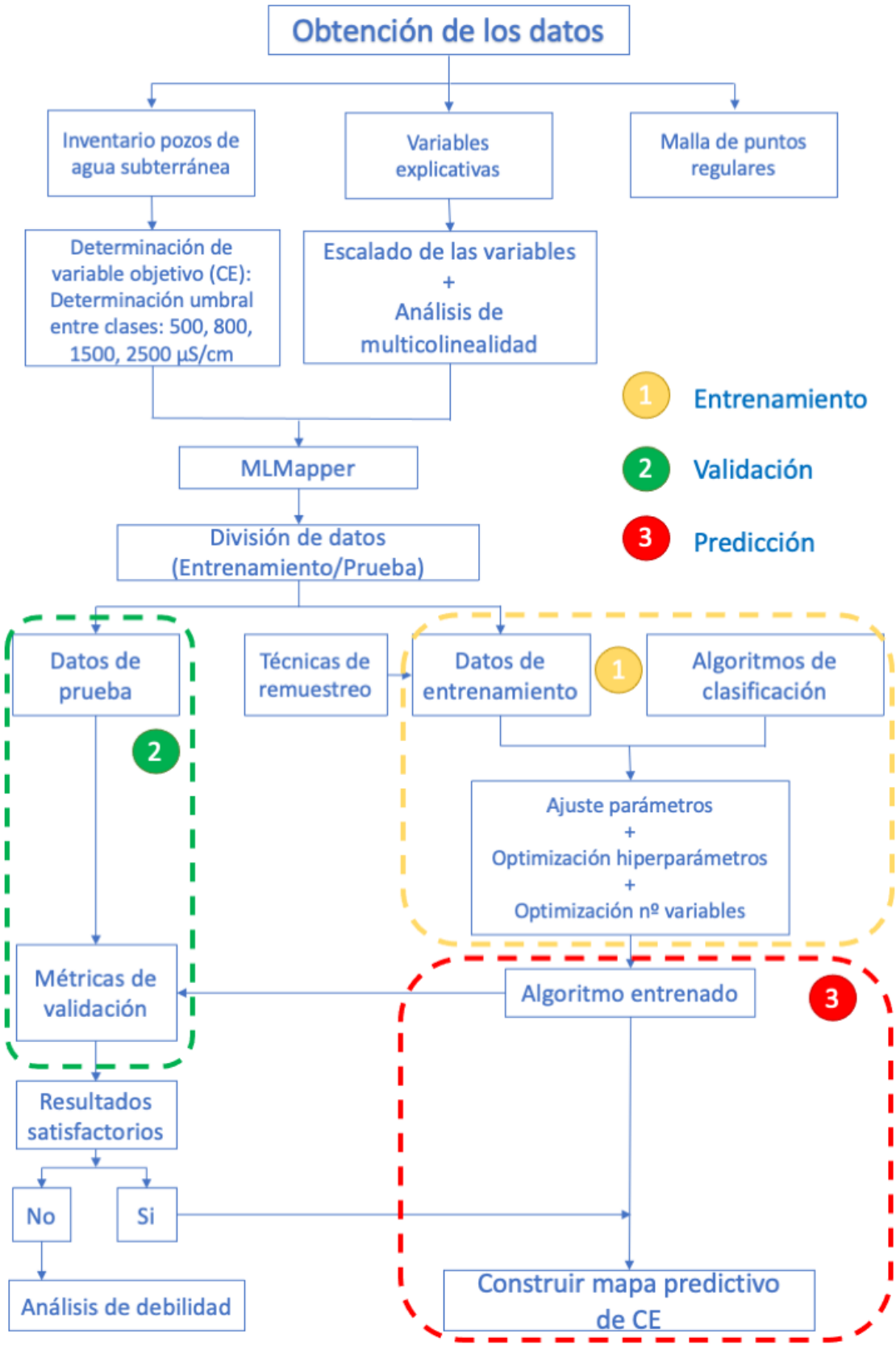

Figura 14: Esquema de funcionamiento de MLMapper v2.0 aplicado a la predicción espacial de CE de las aguas subterráneas (Modificado de Gómez-Escalonilla, 2024).

3.5.1. Preprocesamiento de las variables explicativas

Antes de incorporar los ficheros en el proceso de aprendizaje-predicción, es necesario realizar un preprocesamiento de los datos para su correcta ejecución.

En ocasiones, las variables explicativas tienen rangos de valores numéricos muy diferentes entre ellos y, por lo tanto, pueden tener diferentes impactos o influencias muy dispares atendiendo únicamente al rango y no a las relaciones existentes entre la variable objetivo y las variables explicativas. En ese sentido, los algoritmos podrían otorgar más peso a algunas variables que a otras solo por el rango de valores que contiene (Gómez-Escalonilla, 2024). En este trabajo, por ejemplo, la evapotranspiración presenta valores entre 36,5 y 1024,5 mm/año y, en cambio, el NDVI presenta valores entre -0,33 y 0,57. Esto puede hacer que la evapotranspiración tenga un peso mayor por presentar valores mayores. Esta disparidad entre los valores numéricos de distintas variables explicativas se puede solucionar usando las técnicas de escalado de datos. Dichos métodos consiguen ajustar los valores numéricos de las variables explicativas en un mismo rango. En el presente trabajo se ha utilizado el escalador máximo absoluto (*Maximal absolute scaler - MaxAbs*), el cual ajusta los datos de las variables explicativas usadas durante el entrenamiento dentro del rango [-1,1] y los divide por el valor máximo mayor de cada característica, en caso de haber datos negativos (Pedregosa et al., 2011; Zheng & Casari, 2018).

Otro factor fundamental a tener en cuenta en este tipo de trabajos es el análisis de colinealidad. Es posible que ciertas variables explicativas se encuentren muy correlacionadas entre ellas y presenten una elevada multicolinealidad. Esto puede hacer, al igual que en el caso anterior, que los algoritmos atribuyan un peso extra a una de las variables explicativas o incluso que añadan ruido a los resultados finales (Dormann et al., 2013). Para evaluar esta problemática, en este trabajo se ha utilizado el método del coeficiente de correlación de Pearson, el cual es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente.

Los valores del coeficiente de correlación de Pearson (r) varían entre -1 y 1 y se han representado mediante una matriz de correlación. Cuando los valores de r son positivos existe una correlación directa entre ambas variables explicativas. En cambio, cuando los valores de r son negativos se da una correlación inversa. Además, mediante el valor de r también se puede determinar la fuerza de la correlación. Si los valores de r son cercanos a 0 existe una baja correlación entre las dos variables y si los valores de r se acercan más a ± 1 existe una fuerte correlación. No existe un consenso en la comunidad científica acerca de cuál es el valor máximo de r asumible para excluir problemas de multicolinealidad. Dormann et al. (2013) indican que dos variables son redundantes entre sí cuando el valor de r sea mayor a 0,7 o menor a -0,7. Sin embargo, también hay que tener en cuenta que tipo de relación tienen las

dos variables consideradas redundantes entre sí, ya que puede ser el caso que, aunque exista una gran correlación entre ellas, ambas aporten información valiosa.

3.5.2. Técnicas de remuestreo

En ciertas ocasiones puede ocurrir que una o varias de las clases dentro de un problema de clasificación supervisada se encuentren sobrerrepresentadas en relación con las otras (Lemaître et al., 2017). Esto puede provocar que los algoritmos ignoren cierta información asociada a las clases minoritarias. En la base de datos empleada en este trabajo se ha observado una mayoría de valores de la variable objetivo correspondientes a 0 o negativos, es decir, que no superan los umbrales de conductividad eléctrica establecidos, frente a valores 1 o positivos, que son aquellos que sí superan los umbrales de conductividad eléctrica establecidos. Por tanto, se ha optado por emplear técnicas de sobremuestreo, lo cual consiste en generar nuevas muestras de la clase minoritaria para equilibrar el conjunto de datos, en este caso de los valores correspondientes a la clase positiva. Para el sobremuestreo se han empleado dos técnicas diferentes. La primera técnica se conoce como sobremuestreo sintético minoritario (*Synthetic Minority Oversampling Technique*, SMOTE) (Chawla et al., 2002) y consiste en coger cada muestra de la clase minoritaria y generar muestras en puntos aleatorios sobre las líneas que unen a los vecinos de la clase minoritaria. Por otro lado, la segunda técnica se denomina sobremuestreo sintético adaptativo (*Adaptive Synthetic Sampling approach*, ADASYN), la cual es similar a SMOTE, pero con la diferencia de que además se centra en generar más muestras sintéticas para las muestras de clase minoritaria que tienen vecinos de muestras de clase mayoritaria cerca (He et al., 2008).

3.5.3. Validación con métricas de aprendizaje automático

Las predicciones realizadas por los algoritmos en la etapa de validación tienen que ser evaluadas mediante diferentes métricas. Las métricas de evaluación se definen como una herramienta que mide y evalúa el rendimiento de los algoritmos de clasificación supervisada (Hossin & Sulaiman, 2015). En la fase de validación se ha empleado el 30 % de la base de datos original.

En los enfoques de clasificación binaria, la evaluación de la solución óptima puede analizarse mediante la matriz de confusión. Las columnas de dicha matriz representan la clase real y las filas representan la clase predicha por el modelo (Figura 15). Cuando la clase real y la clase predicha coinciden, los puntos serán definidos como verdaderos positivos o verdaderos negativos, y cuando la clase real y la clase predicha no coincidan como falsos positivos o falsos negativos. Dicho de otro modo, TP y TN indican el número de puntos positivos y negativos clasificados correctamente. Por su parte, FP y FN indican el número de casos

positivos y negativos clasificados de manera errónea, respectivamente (Hossin & Sulaiman, 2015).

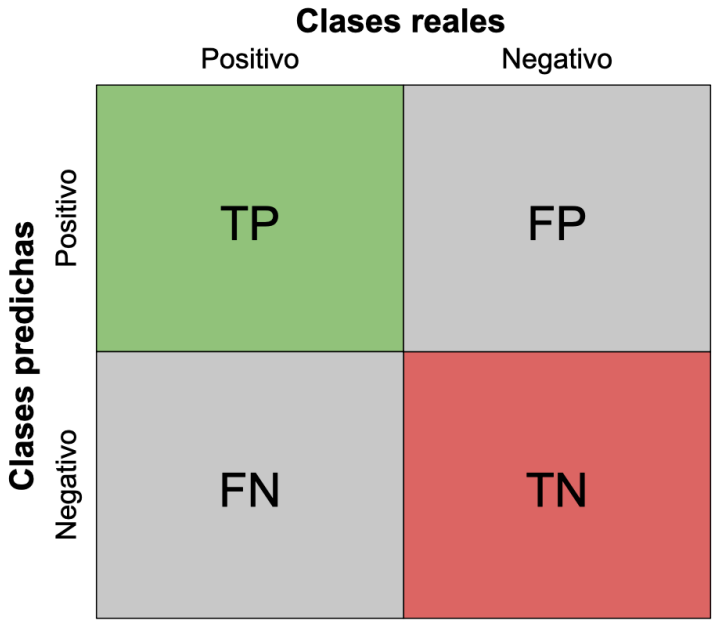

Figura 15: Esquema de una matriz de confusión, donde TP=verdadero positivo, FP= falso positivo, FN=falso negativo y TN=verdadero negativo (Modificado de Gómez-Escalonilla, 2024).

A partir de la información extraída de la matriz de confusión se pueden generar varias métricas con diferentes enfoques de evaluación. Las métricas empleadas en este trabajo son las siguientes:

En primer lugar, el *Test score* o la puntuación de la prueba, que hace referencia al número de predicciones correctas sobre el número total de intentos (Gómez-Escalonilla, 2024). Esta métrica oscila entre el 0 y el 1, siendo los valores cercanos al 0 indicativos de una baja capacidad predictiva y los valores cercanos al 1 indicativos de una alta capacidad predictiva.

$$Test\ score = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Por otro lado, la precisión se puede calcular tanto para las clases positivas como para las clases negativas. En el caso de las clases positivas, se calcula como los verdaderos positivos sobre todos los puntos que el modelo ha predicho como positivos (Gómez-Escalonilla, 2024). Por el contrario, para las clases negativas se calcula como los verdaderos negativos sobre los puntos que el modelo ha predicho como negativos. Como en la métrica anteriormente descrita, el valor 0 es la puntuación más baja y el valor 1 la más alta.

$$\text{Precisión (positiva)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\text{Precisión (negativa)} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

El *Recall* o sensibilidad también se puede calcular para ambas clases, positiva y negativa por separado. Para el caso de las positivas se calcula la fracción de clases positivas que son correctamente clasificadas (Hossin & Sulaiman, 2015), es decir, la relación entre los

verdaderos positivos y todos los positivos reales. En cambio, para las clases negativas se calcula la relación entre los verdaderos negativos y los negativos reales. En este caso también el valor 0 es el más bajo y el valor 1 el más alto.

$$Recall\ (\text{positiva}) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$Recall\ (\text{negativa}) = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

El *F1-score* o la puntuación F1 calcula la media harmónica entre la precisión y el *recall* (Hossin & Sulaiman, 2015). Se puede calcular tanto el *F1-score* para las clases positivas como el *F1-score* para las clases negativas. Sus valores oscilan entre 0 y 1, siendo el 0 el valor más bajo y el 1 el valor más alto.

$$F1\ score\ (\text{positiva}) = 2 \times \frac{\text{Precisión (positiva)} \times\ Recall\ (\text{positiva})}{\text{Precisión (positiva)} + Recall\ (\text{positiva})}$$

$$F1\ score\ (\text{negativa}) = 2 \times \frac{\text{Precisión (negativa)} \times\ Recall\ (\text{negativa})}{\text{Precisión (negativa)} + Recall\ (\text{negativa})}$$

Por último, una de las métricas más populares en los enfoques de clasificación binaria supervisada, es el área bajo la curva de características operativas del receptor (*Area Under Receiver Operating Characteristic curve, AUC ROC*). Dicha curva representa dos métricas: la tasa de verdaderos positivos (*True Positive Rate, TPR*) en el eje Y y la tasa de falsos positivos (*False Positive Rate, FPR*) en el eje X, las cuales se calculan mediante las siguientes fórmulas.

$$TPR = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

El AUC corresponde al área que queda debajo de la curva ROC de un determinado modelo. Un AUC de 0.5 indica que no existe ninguna discriminación entre ambas clases y los valores más cercanos a 1 indican que el modelo tiene un buen rendimiento, es decir, de 0.7 a 0.8 se considera aceptable, de 0.8 a 0.9 se considera excelente y por encima de 0.9 se considera excepcional (Hosmer & Lemeshow, 2000).

3.5.4. Técnicas para mejorar la interpretabilidad de los algoritmos

Con el objetivo de mejorar la interpretabilidad de las predicciones realizadas por los diferentes modelos, en este trabajo se ha utilizado el método de la importancia de las variables o *Feature importance*. Esta técnica calcula una puntuación para cada una de las variables y un modelo dado, es decir, calcula el efecto o influencia que tiene cada variable explicativa en un determinado modelo a la hora de tomar decisiones y predecir una variable objetivo.

Este método se representa en un diagrama donde en el eje Y se indican las diferentes variables explicativas y en el eje X la fracción esperada de las instancias a las que contribuyen dichas variables. Las variables situadas en la parte superior del árbol tienen mayor valor en el

eje X y, por tanto, mayor efecto en la decisión de la predicción final (Gómez-Escalonilla, 2024). De este modo, el método *Feature importance* indica las variables explicativas que más contribuyen a la correcta predicción de la variable objetivo (Gómez-Escalonilla, 2024). Este método solo se puede aplicar en los modelos pertenecientes a la familia de algoritmos "basados en árboles".

3.5.5. Cartografía predictiva y estimación de población en riesgo

Con los algoritmos con mejor rendimiento se han realizado cartografías predictivas calculando la media pixel por pixel de la probabilidad de exceder un determinado umbral de conductividad eléctrica. Además, se ha realizado un análisis cualitativo de la estimación de la población en riesgo por consumo de las aguas subterráneas con elevada salinidad. Para ello, se ha obtenido un archivo ráster de densidad de población (habitantes/hectárea) correspondiente al año 2015 y con una resolución espacial de 100x100 metros (WorldPop, 2017).

4. RESULTADOS Y DISCUSIÓN

4.1. Análisis de multicolinealidad

La Figura 16 representa la matriz de correlación correspondiente a las 18 variables explicativas empleadas en este trabajo. Se observan diferentes valores del coeficiente de correlación Pearson representadas con diferentes colores, desde el valor -1 con color rojo oscuro (correlación inversa) hasta el valor 1 con color azul oscuro (correlación directa).

En este trabajo, se puede ver una fuerte correlación tanto directa (> 0.7) como inversa (< -0,7) en varios pares de variables explicativas. Se ha observado una fuerte correlación directa entre "Profundidad del basamento impermeable" y "Espesor saturado", entre "Precipitación" y "Evapotranspiración real", entre "Recarga" y "Evapotranspiración real" y entre "Recarga" y "Precipitación". A su vez, se ha observado una fuerte correlación inversa entre "NDWI" y "NDVI" y entre "Temperatura" y "Elevación".

Estas correlaciones se esperaba que ocurrieran desde un principio. En el primer caso, el espesor saturado fue obtenido a partir de la profundidad del nivel freático y la profundidad del basamento impermeable, por tanto, se esperaba una fuerte correlación entre ambas variables. En el caso de las variables climáticas, la precipitación condiciona de manera directa la tasa de evapotranspiración real y, al mismo tiempo, la recarga está altamente condicionada por la precipitación y evapotranspiración real, siendo la precipitación la principal fuente de agua de recarga. En el caso de las correlaciones inversas, por lo general la temperatura disminuye a medida que aumenta la elevación del terreno, por lo que se podía prever la correlación inversa entre ambos factores. Respecto a los índices NDWI y el NDVI, la correlación inversa observada se explica por qué los valores de NDWI cercanos al 1 indican la presencia de una superficie de agua, por ejemplo, un lago o un río, lo cual es incompatible con la existencia de grandes cantidades de vegetación, por ejemplo, un bosque, como indican los valores de NDVI cercanos al 1.

A pesar de haber observado varias correlaciones fuertes entre las diferentes variables explicativas mencionadas anteriormente, ninguna de estas variables se ha considerado redundante debido a su valiosa aportación al resultado final y, por tanto, no se ha eliminado ninguna variable explicativa.

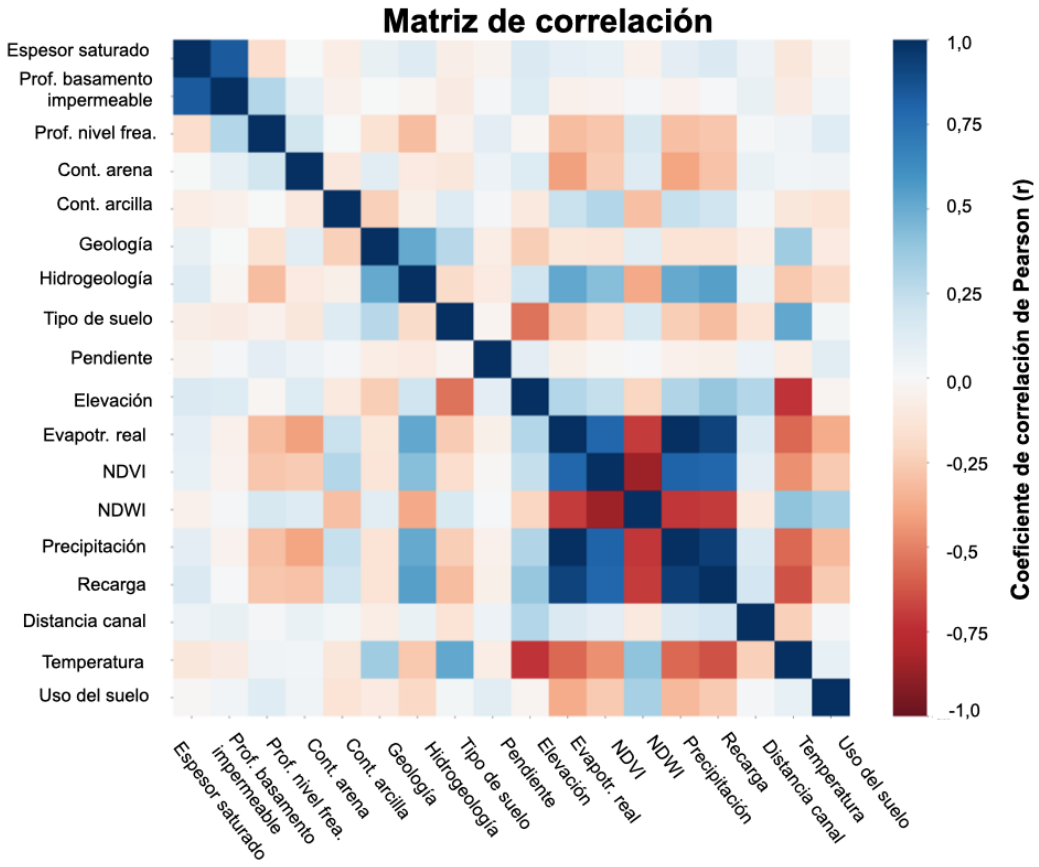

Figura 16: Matriz de correlación con las variables explicativas utilizadas.

4.2. Evaluación de los algoritmos

A continuación, se muestran los valores de diferentes métricas de evaluación obtenidos por los algoritmos con los mejores resultados en cada uno de los tres escenarios (sin remuestreo, con remuestreo ADASYN y con remuestreo SMOTE) y para los 4 umbrales de CE evaluados, es decir, 500 µS/cm, 800 µS/cm, 1500 µS/cm y 2500 µS/cm (Tablas 3, 4, 5 y 6).

Para el caso del umbral de CE de 500 µS/cm, los algoritmos que han obtenido mejores resultados en las métricas de validación son *Extra Trees Classifier* (ETC), *Random Forest Classifier* (RFC), *Gradient Boosting Classifier* (RBC) y *Linear Discriminant Analysis* (LDA) (Tabla 3). Respecto a los escenarios de aplicación de técnicas de remuestreo, se han obtenido resultados ligeramente mejores cuando no se aplicaban estas técnicas sobre la clase minoritaria. En todos los casos el *test score* supera el valor de 0,72 y el AUC supera el valor de 0,83. En cuanto a las demás métricas de validación, se observan valores más altos en las métricas correspondientes a las clases negativas (0) en comparación con las correspondientes a las clases positivas (1). Los valores de *F1-score* (0) oscilan entre 0,79 y 0,89, mientras que los valores de *F1-score* (1) oscilan entre 0,6 y 0,66. Esto indica que los algoritmos son capaces de distinguir entre las clases negativas y positivas de manera precisa, aunque cabe destacar que tienen mejor rendimiento a la hora de predecir la clase negativa. Puesto que el umbral de CE de 500 µS/cm es relativamente bajo, no existe una elevada

disparidad entre el número de puntos pertenecientes a la clase negativa y a la clase positiva en los datos de entrada. Esto podría explicar por qué los resultados de las métricas de validación para los tres escenarios (sin remuestreo, con remuestreo ADASYN y con remuestreo SMOTE) son similares e incluso mejores en el caso de sin remuestreo. Cabe destacar que, en este caso, la mayoría de los algoritmos con buenos resultados pertenecen a la familia de algoritmos "basados en árboles": *Extra Trees Classifier*, *Random Forest Classifier* y *Gradient Boosting Classifier.* Sin embargo, para elaborar la cartografía predictiva se han empleado los tres algoritmos con mejor rendimiento, es decir, ETC, RFC y LDA.

Tabla 3: Valores de métricas de validación para los algoritmos con mejor rendimiento en el caso del umbral de CE= 500 µS/cm.

CE= 500 µs/cm	Algoritmo	Test score	Precisión (0)	Precisión (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)	AUC	Media
Sin remuestreo	Extra Trees	0,828	0,870	0,680	0,900	0,620	0,890	0,650	0,860	0,787
	Random Forest	0,825	0,870	0,670	0,890	0,620	0,880	0,650	0,864	0,784
	Linear Discriminant	0,828	0,870	0,690	0,910	0,600	0,890	0,640	0,833	0,783
Remuestreo ADASYN	Gradient Boosting	0,800	0,880	0,600	0,850	0,660	0,860	0,630	0,844	0,766
	Extra Trees	0,754	0,930	0,510	0,720	0,840	0,810	0,640	0,858	0,758
	Random Forest	0,754	0,920	0,510	0,730	0,820	0,810	0,630	0,854	0,753
Remuestreo SMOTE	Extra Trees	0,793	0,910	0,570	0,800	0,770	0,850	0,660	0,859	0,776
	Gradient Boosting	0,814	0,880	0,630	0,870	0,650	0,870	0,640	0,849	0,775
	Random Forest	0,778	0,920	0,550	0,770	0,790	0,840	0,650	0,857	0,769

En el escenario que emplea un umbral de CE de 800 µS/cm para discriminar entre puntos positivos y negativos, los algoritmos con mejor rendimiento son *Gradient Boosting Classifier* (GBC), *Random Forest Classifier* (RFC), *Gaussian Naive Bayes* (GNB), *Quadratic Discriminant Analysis* (QDA), *Linear Discriminant Analysis* (LDA) y *Decision Tree Classifier* (DTC) (Tabla 4). En este caso se obtienen mejores resultados de las métricas de evaluación en los dos escenarios donde se ha realizado un sobremuestreo (ADASYN y SMOTE). Los valores de *Test Score* y AUC son ligeramente más elevados en el escenario sin remuestreo aunque estas métricas superan los valores 0,75 y 0,8 en todos los casos, respectivamente. Los valores de las métricas correspondientes a las clases positivas, Precisión (1), *Recall* (1) y *F1-score* (1), son notablemente más bajos que en el caso del umbral de CE de 500 µS/cm. Sin embargo, dichos valores son más altos en el caso de los escenarios donde se ha realizado un sobremuestreo en comparación con el escenario donde no se ha realizado un remuestreo lo que evidencia que, bajo el escenario sin remuestreo tiende a sobreestimar la clase mayoritaria.

A medida que aumenta el umbral de CE, los algoritmos encuentran más dificultades a la hora de predecir clases positivas (1) y, por tanto, presentan mejores resultados en los escenarios donde se realiza un sobremuestreo de la clase minoritaria, en este caso la clase positiva. Esto

se debe al menor número de muestras en la clase positiva conforme aumenta el umbral discriminatorio. Finalmente, bajo el escenario que emplea un umbral de 800 µS/cm, la técnica de remuestreo con mejor resultado es el SMOTE y los algoritmos con mejor rendimiento que serán empleados para elaborar la cartografía predictiva son GBC, LDA y RFC.

Tabla 4: Valores de métricas de validación para los algoritmos con mejor rendimiento en el caso del umbral de CE= 800 µS/cm.

CE= 800 µs/cm	Algoritmo	Test score	Precisión (0)	Precisión (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)	AUC	Media
Sin remuestreo	Gradient Boosting	0,889	0,920	0,510	0,960	0,360	0,940	0,420	0,857	0,732
	Quadratic Discriminant	0,812	0,960	0,340	0,830	0,710	0,890	0,460	0,840	0,730
	Gaussian NB	0,760	0,970	0,300	0,750	0,840	0,850	0,440	0,841	0,719
Remuestreo ADASYN	Random Forest	0,804	0,970	0,340	0,810	0,780	0,880	0,470	0,854	0,738
	Decission Tree	0,789	0,970	0,320	0,790	0,780	0,870	0,450	0,826	0,724
	Gradient Boosting	0,798	0,960	0,320	0,810	0,730	0,880	0,450	0,840	0,723
Remuestreo SMOTE	Gradient Boosting	0,840	0,950	0,370	0,870	0,620	0,910	0,460	0,851	0,734
	Linear Discriminant	0,763	0,980	0,300	0,750	0,860	0,850	0,450	0,846	0,725
	Random Forest	0,759	0,980	0,300	0,750	0,860	0,850	0,440	0,842	0,723

En el caso de emplear el umbral de CE de 1500 µS/cm, los algoritmos con mejores resultados en las métricas de evaluación son *Quadratic Discriminant Analysis*, *Extra Trees Classifier*, *K Neighbors Classifier*, *Linear Discriminant Analysis* y *Gaussian Naive Bayes* (Tabla 5). Los valores de *Test score* oscilan entre 0,71 y 0,95 y los valores de AUC varían entre 0,80 y 0,85. Los valores de las demás métricas correspondientes a las clases negativas (0), Precisión (0), *Recall* (0) y *F1-score* (0), muestran buenos resultados. Sin embargo, las métricas de evaluación correspondientes a las clases positivas (1), Precisión (1) y *F1-score* (1), muestran resultados muy bajos, no llegando a superar los valores 0,14 y 0,24, respectivamente. Además, no se observa ninguna mejora en cuanto a los valores de las métricas correspondientes a las clases positivas (1) en los escenarios donde se ha realizado un remuestreo (ADASYN y SMOTE). A tenor de los resultados obtenidos, es evidente que el número de puntos positivos no es lo suficientemente grande y representativo como para encontrar asociaciones significativas entre variables explicativas y variable objetivo.

Tabla 5: Valores de métricas de validación para los algoritmos con mejor rendimiento en el caso del umbral de CE= 1500 µS/cm.

CE= 1500 µs/cm	Algoritmo	Test score	Precisión (0)	Precisión (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)	AUC	Media
Sin remuestreo	Quadratic Discriminant	0,850	0,990	0,140	0,860	0,650	0,920	0,230	0,824	0,683
	Gaussian NB	0,747	0,990	0,100	0,750	0,770	0,850	0,180	0,828	0,652
	Linear Discriminant	0,950	0,970	0,120	0,980	0,070	0,970	0,090	0,807	0,620
Remuestreo ADASYN	K Neighbors	0,840	0,980	0,130	0,850	0,630	0,910	0,220	0,834	0,674
	Extra Trees	0,781	0,990	0,110	0,780	0,770	0,870	0,200	0,840	0,668
	Quadratic Discriminant	0,823	0,980	0,120	0,830	0,630	0,900	0,200	0,806	0,661
Remuestreo SMOTE	K Neighbors	0,827	0,990	0,140	0,830	0,790	0,900	0,240	0,831	0,693
	Extra Trees	0,720	1,000	0,100	0,710	0,910	0,830	0,180	0,853	0,663
	Quadratic Discriminant	0,823	0,980	0,120	0,830	0,630	0,900	0,200	0,808	0,661

Por último, en lo que corresponde al umbral de CE de 2500 µS/cm, los algoritmos con mejor rendimiento son *K Neighbors Classifier*, *Ada Boost Classifier*, *Extra Trees Classifier*, *Gaussian Naive Bayes*, *Random Forest Classifier* y *Linear* Support *Vector Classifier* (Tabla 6). Se observan unos valores de las métricas similares a los valores correspondientes al umbral de CE de 1500 µS/cm. El *Test score* y el AUC muestran valores altos, superando 0,74 y 0,71 en todos los casos, respectivamente. Además, los valores de Precisión y *F1-score* correspondientes a la clase negativa (0) indican que los algoritmos han obtenido buenos resultados, llegando en algunos casos al valor de 1. Por el contrario, e igual que en el caso del umbral de CE de 1500 µS/cm, los valores de Precisión y *F1-score* correspondientes a la clase positiva (1) son muy bajos, llegando a 0,14 en el mejor de los casos. A su vez, se observa una ligera mejora en los resultados de los escenarios donde se ha realizado un remuestreo.

Dados los resultados obtenidos, para los umbrales de CE de 1500 µS/cm y 2500 µS/cm, se puede concluir que ningún algoritmo es capaz de predecir la clase positiva de manera adecuada y que, además, la aplicación de las dos técnicas de remuestreo no ha resultado exitosa. Por ello, los modelos obtenidos con estos umbrales no han sido empleados para elaborar las cartografías y en análisis posterior.

Tabla 6: Valores de métricas de validación para los algoritmos con mejor rendimiento en el caso del umbral de CE= 2500 µS/cm.

CE= 2500 µs/cm	Algoritmo	Test score	Precisión (0)	Precisión (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)	AUC	Media
Sin remuestreo	Gaussian NB	0,789	0,990	0,030	0,790	0,640	0,880	0,060	0,736	0,614
	Random Forest	0,989	0,990	0,000	1,000	0,000	0,990	0,000	0,847	0,602
	Extra Trees	0,989	0,990	0,000	1,000	0,000	0,990	0,000	0,832	0,600
Remuestreo ADASYN	K Neighbors	0,892	0,990	0,050	0,900	0,500	0,940	0,100	0,721	0,637
	Ada Boost	0,892	0,990	0,050	0,900	0,430	0,940	0,080	0,796	0,635
	Extra Trees	0,746	1,000	0,030	0,750	0,790	0,850	0,070	0,778	0,627
Remuestreo SMOTE	K Neighbors	0,889	0,990	0,050	0,890	0,500	0,940	0,090	0,718	0,633
	Ada Boost	0,942	0,990	0,060	0,950	0,290	0,970	0,100	0,754	0,632
	Support Vector	0,985	0,990	0,140	1,000	0,070	0,990	0,100	0,751	0,628

4.3. Importancia de las variables explicativas

La Figura 17 representa la importancia de las variables para dos de los algoritmos con el mejor resultado, el *Extra Trees Classifier* (ETC) y el *Random Forest* Classifier (RFC), cuando se utiliza el umbral de CE de 500 µS/cm y para el escenario sin remuestreo. La gráfica correspondiente al algoritmo ETC muestra que la temperatura, la recarga de agua subterránea, la geología, la evapotranspiración real, la hidrogeología y la precipitación son las variables explicativas más importantes a la hora de condicionar la conductividad eléctrica de las aguas subterráneas. La importancia combinada de las 6 variables explicativas mencionadas anteriormente es de aproximadamente 0,815. Sin embargo, se observa que la importancia relativa de las variables explicativas varía dependiendo del algoritmo. Comparando los resultados de la importancia de las variables calculada para el algoritmo ETC con el siguiente algoritmo con el mejor rendimiento para el umbral de CE de 500 µS/cm, *Random Forest Classifier* (RFC), se puede observar que ambos coinciden en la temperatura, la recarga, la hidrogeología y la evapotranspiración real como algunas de las variables más importantes. En el caso del RFC, las variables explicativas más importantes son la recarga, la temperatura, la hidrogeología, la elevación del terreno, la profundidad del nivel freático y la evapotranspiración real. Estas 6 variables explicativas suman una importancia aproximada de 0,75.

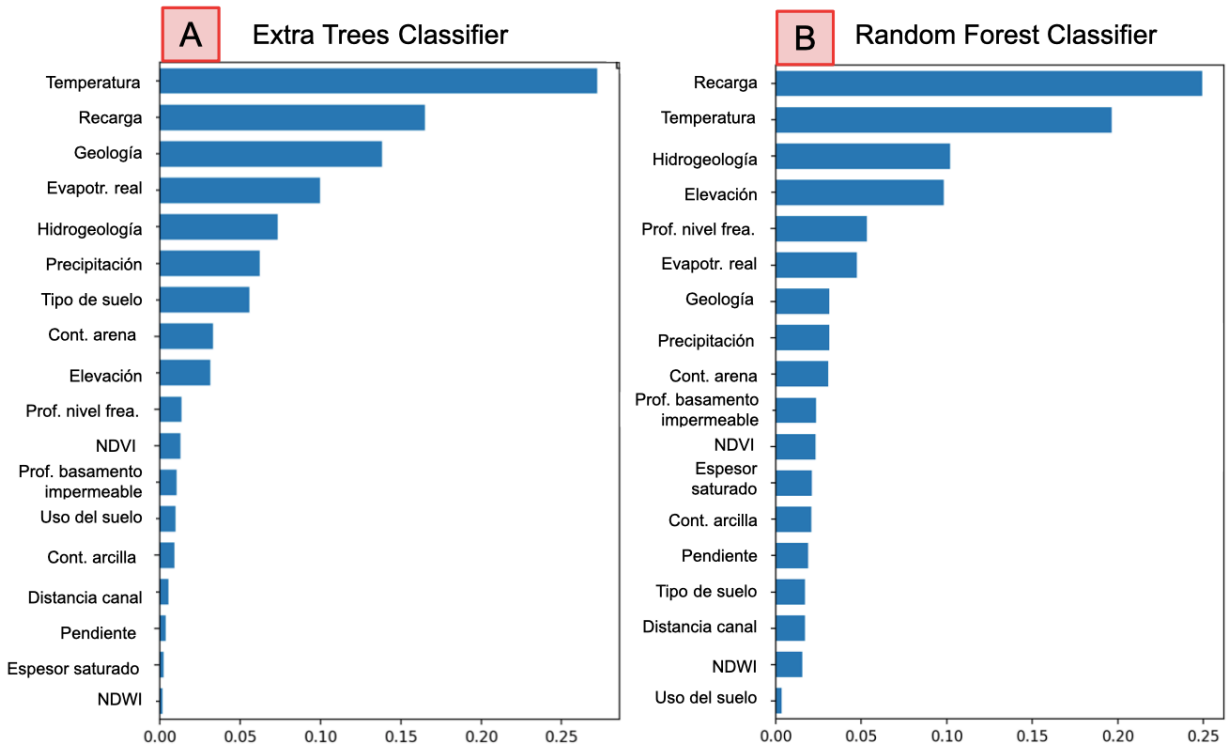

Figura 17: La importancia de las variables calculada para los dos algoritmos con mejor rendimiento para el umbral de CE de 500 µS/cm: A) El algoritmo ETC. B) El algoritmo RFC. El sumatorio de la importancia de las variables explicativas es igual a 1.

Por otro lado, la Figura 18 muestra la importancia de las variables explicativas para los dos algoritmos "basados en árboles" con mejor rendimiento, *Gradient Boosting Classifier* (GBC) y *Random Forest Classifier* (RFC), para el caso del umbral de CE de 800 µS/cm y bajo el escenario donde se emplea la técnica de remuestreo SMOTE. Como se puede observar en la gráfica correspondiente al algoritmo GBC las variables más importantes a la hora de controlar la salinidad de las aguas subterráneas son la evapotranspiración real, la precipitación, la recarga de agua subterránea, el tipo de suelo, la profundidad del nivel freático y el uso del suelo. Estas variables explicativas suman una importancia combinada de aproximadamente 0,725. En cambio, el segundo algoritmo con mejor resultado para el umbral de CE de 800 µS/cm, el algoritmo RFC, considera como variables más importantes a la evapotranspiración real, la recarga de agua subterránea, la precipitación, el tipo de suelo, la hidrogeología y la elevación. La importancia combinada de estas variables es de 0,9. Ambos algoritmos coinciden en que la evapotranspiración real, la recarga, la precipitación y el tipo de suelo son algunas de las variables explicativas más importantes.

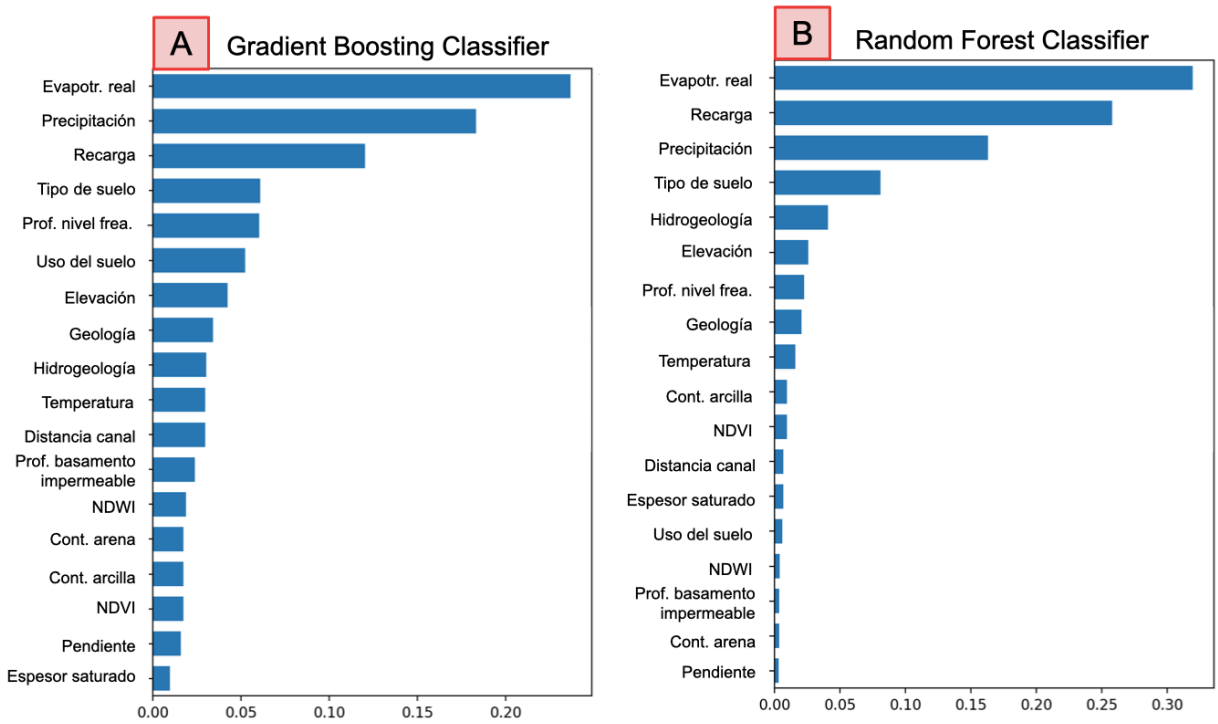

Figura 18: La importancia de las variables calculada para los dos algoritmos con mejor rendimiento para el umbral de CE de 800 µS/cm: A) El algoritmo GBC. B) El algoritmo RFC. El sumatorio de la importancia de las variables explicativas es igual a 1.

Estos resultados coinciden con los obtenidos por otros autores como Araya et al. (2023), los cuales obtuvieron, empleando un enfoque similar, la precipitación, la recarga, la elevación del terreno, el coeficiente de *Priestley Taylor* (el cual relaciona la evapotranspiración real y la evapotranspiración potencial), la distancia a la costa y la geología, como variables explicativas importantes en este orden. A su vez, Sahour et al. (2020) obtuvieron como variables explicativas importantes la transmisividad, la precipitación, el nivel freático, la distancia a la costa y la elevación del terreno.

4.4. Cartografía predictiva

A continuación, se presentan los mapas correspondientes a la media de la probabilidad (de 0 a 1) de exceder un determinado umbral de conductividad eléctrica predicha por los algoritmos con los mejores resultados. La Figura 19 hace referencia a la probabilidad de exceder el umbral de CE de 500 µS/cm para los algoritmos ETC, RFC y LDA en el escenario donde no se ha realizado un remuestreo. Por su parte, la Figura 20 hace referencia a la probabilidad de exceder el umbral de CE de 800 µS/cm para los algoritmos GBC, LDA y RFC en el escenario en el que se ha empleado la técnica de remuestreo SMOTE.

En la Figura 19 se puede observar que los valores más bajos de probabilidad de exceder el umbral de CE de 500 µS/cm se encuentran en su mayoría en la zona suroccidental del área de estudio. Por el contrario, los valores más elevados de probabilidad se sitúan al noroeste, en la frontera con Mauritania, y al sureste de la zona de estudio. Asimismo, se puede observar

una disminución gradual de los valores de probabilidad de norte a sur. Comparando con la distribución espacial de los valores de probabilidad de exceder el umbral de CE de 800 µS/cm (Figura 20) se puede observar una clara similitud. De la misma manera, en la Figura 20 se puede ver que los valores de probabilidad más bajos se encuentran al suroeste de la zona de estudio y los valores de probabilidad más altos se sitúan sobre todo al noroeste y al sureste. Cabe destacar que comparando con las probabilidades de exceder el umbral de CE de 500 µS/cm, las probabilidades de exceder el umbral de CE de 800 µS/cm son todavía más bajos al suroeste y más altos al noroeste y al sureste de la zona de estudio.

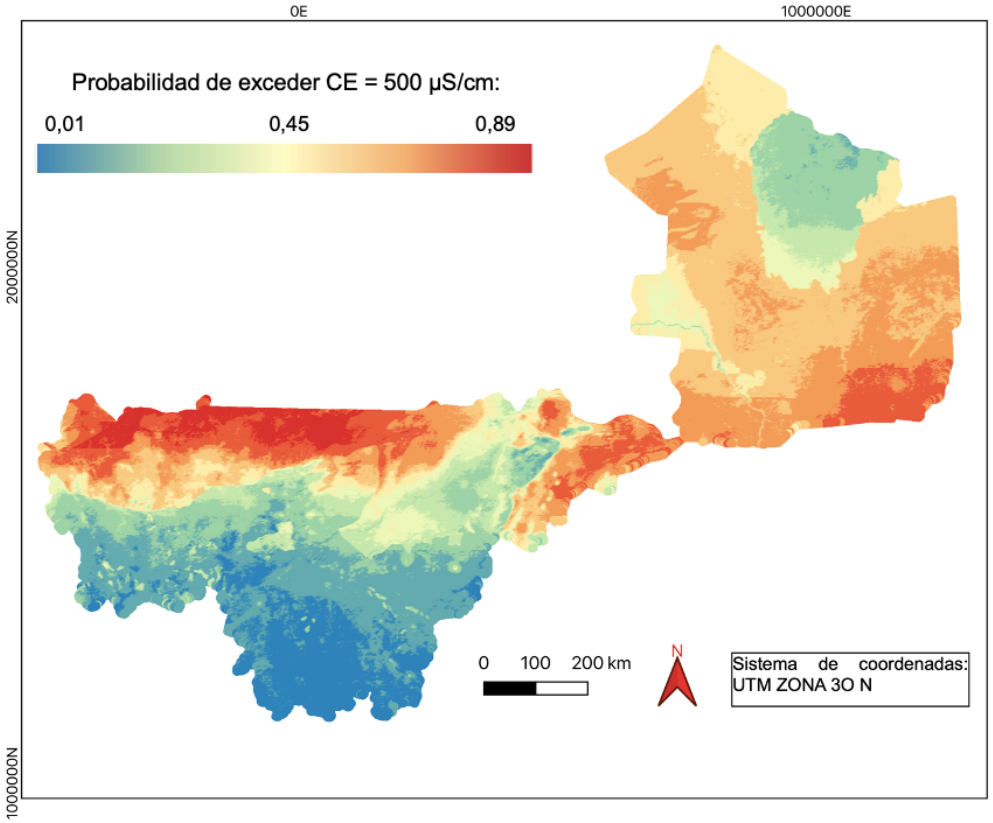

Figura 19: Mapa predictivo de la probabilidad del agua subterránea de exceder el umbral de conductividad eléctrica de 500 µS/cm.

Las zonas con valores de probabilidad más bajos coinciden con las zonas con valores de temperatura más bajos, valores de precipitación más altos y valores de recarga más elevados. En cambio, se ve que las zonas con valores de probabilidad más altos corresponden a las áreas con valores más altos de temperatura. Esto indica que las altas temperaturas se relacionan con un incremento de la tasa de evapotranspiración y esto da como resultado un aumento de la salinidad de las aguas subterráneas. A su vez, un incremento de la precipitación tiene un impacto inverso en la salinidad de las aguas subterráneas, debido a que la precipitación es la principal fuente de recarga de agua de los acuíferos y puede disminuir la salinidad al diluir los iones disueltos en el agua subterránea. Además, se ve una clara correlación entre la elevación del terreno y la probabilidad de exceder el umbral de CE. En las zonas montañosas cercanas a la frontera con Argelia (Adras des Iforas) y cerca del monte

Hombori Tondo, donde se encuentran los puntos más altos de la República de Mali, se observan valores de probabilidad relativamente más bajos en comparación con los valores de probabilidad de la zona circundante. Esto indica que, según los algoritmos de inteligencia artificial, la topografía está inversamente relacionada con la salinidad de las aguas subterráneas, ya que condiciona la profundidad del nivel freático y esto, a su vez, controla la tasa de evaporación y la salinidad del agua subterránea.

Los resultados obtenidos en este trabajo coinciden en gran medida con la cartografía de Díaz-Alcaide et al. (2017). Estos autores calculan la media aritmética de la conductividad eléctrica de las aguas subterráneas de Mali a escala de comuna. Otros estudios que emplean un enfoque similar pero realizados en diferentes contextos geográficos coinciden en que los valores más bajos de CE se encuentran en las zonas con mayor altitud (Araya et al., 2023; Akter et al., 2021; Mosavi et al., 2020) y con mayor precipitación (Sahour et al., 2020).

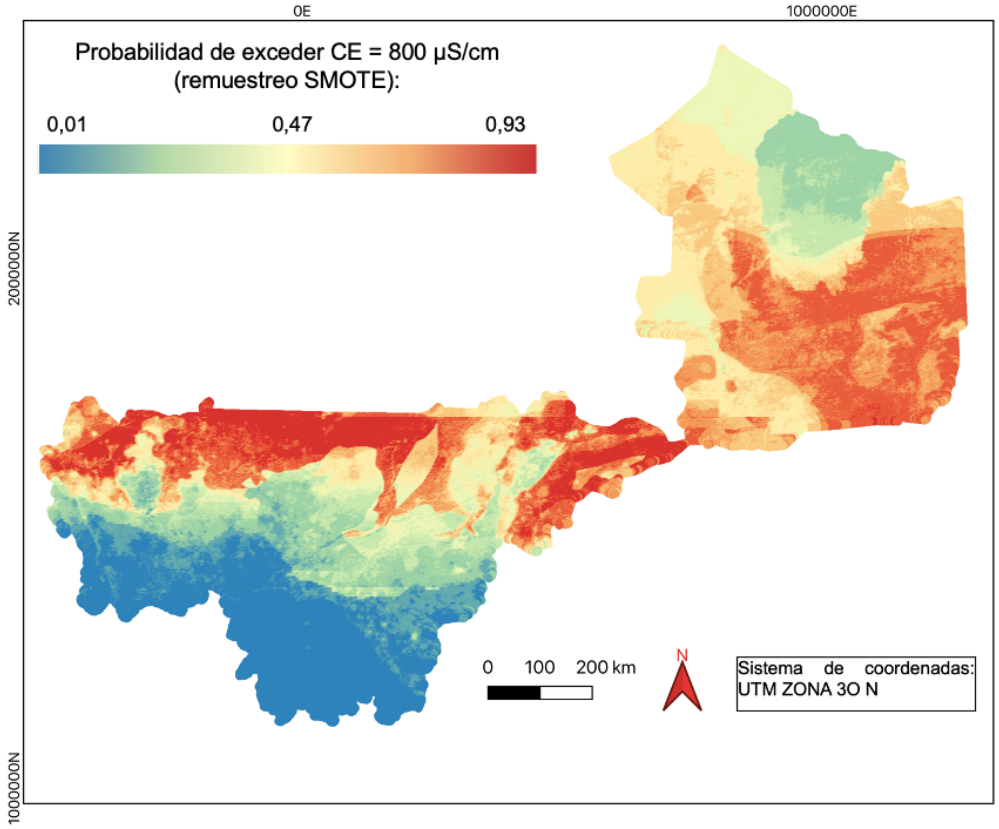

Figura 20: Mapa predictivo de la probabilidad del agua subterránea de exceder el umbral de conductividad eléctrica de 800 µS/cm.

4.5. Limitaciones

Los enfoques de aprendizaje automático, también denominados métodos basados en datos (del inglés, *data-driven methods*), dependen, como su propio nombre indica, de la calidad y cantidad de datos disponibles (Gómez-Escalonilla, 2024).

En este trabajo, la base de datos empleada y proporcionada por la Dirección Nacional de Hidráulica de Mali (Direction Nationale de l'Hydraulique, 2010) incluye datos de más de 21.000

pozos y sondeos distribuidos a lo largo de 8.000 asentamientos de la República de Mali. Sin embargo, la información de todos los sondeos se encuentra referida a la localidad o asentamiento, es decir, no se dispone de la localización de los pozos y sondeos de manera individual. La disponibilidad de esta información permitiría obtener resultados más robustos, puesto que se multiplicaría por tres el número de puntos disponibles para las tareas de aprendizaje y validación. Por otro lado, los datos de conductividad eléctrica disponibles se corresponden a la media de los datos disponibles para una aldea lo que supone una evidente limitación.

Las variables explicativas, por su parte, también conllevan asociadas una serie de limitaciones. En primer lugar, en este trabajo se han empleado 18 variables explicativas, aunque suponen un número elevado de covariables, existe la posibilidad de que no se hayan contemplado factores que muestren una estrecha relación con la salinidad de las aguas subterráneas. Por otro lado, las variables explicativas han sido obtenidas de diferentes fuentes, por lo tanto, también presentan diferencias en cuanto a la resolución espacial. Por ello, las cartografías predictivas presentarán la limitación de adaptarse a la resolución espacial de los datos de partida. Pese a ello, en este trabajo se ha obtenido una cartografía predictiva de la conductividad eléctrica a una escala de 1.000 m para un área total de más de 750.000 km^2, lo que supone una resolución lo suficientemente precisa para la extensión total que abarca el estudio.

4.6. Estimación de población en riesgo

La Figura 21 representa el mapa de densidad de población de la República de Mali, a excepción de la región de Tombouctou, correspondiente al año 2015. Para realizar el análisis de la población en riesgo por consumo de agua subterránea con elevada salinidad se ha utilizado el mapa predictivo de la probabilidad de exceder el umbral de CE de 800 µS/cm ya que dicho umbral corresponde al estándar de agua potable de buena calidad (WHO, 2011). Se puede observar que la densidad de población del país de Mali es relativamente baja y que la zona suroeste está más poblada que el resto del país. En este sector se encuentran la mayoría de los núcleos urbanos importantes de Mali, como Bamako, Sikasso, Koutiala y Segou, entre otros. Por su parte, en la región oriental del área de estudio se encuentran varios núcleos urbanos importantes a orillas del río Níger, como Bourem, Gao y Ansongo. La ubicación de estos núcleos urbanos está condicionada por factores climáticos y geográficos. La mayoría de los núcleos urbanos importantes se sitúan a orillas del río Níger y las zonas de sabana, donde las temperaturas son relativamente más bajas que en el resto del país y las precipitaciones más elevadas.

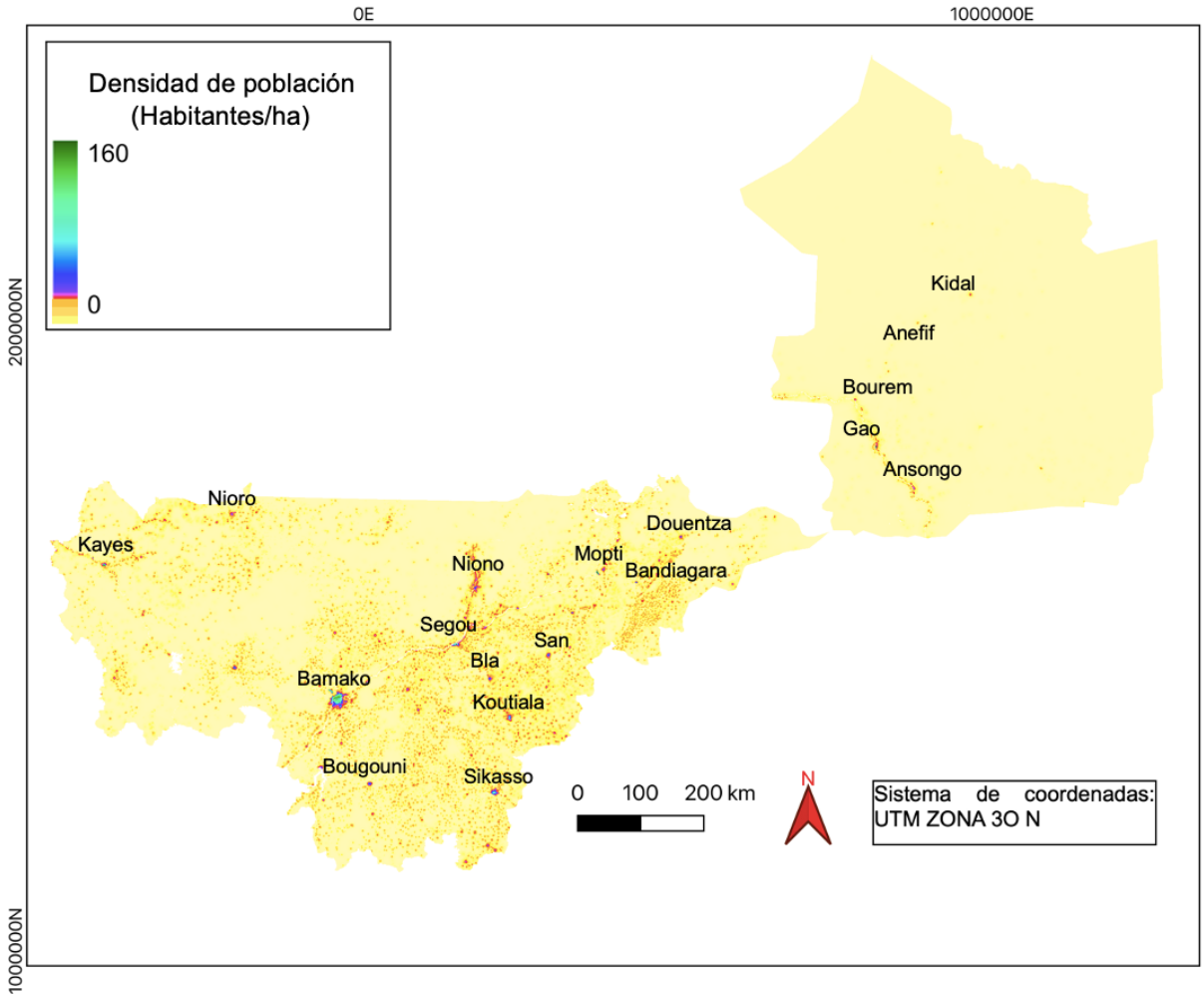

Figura 21: Mapa de la densidad de población de la zona de estudio correspondiente al año 2015 y algunos de los núcleos urbanos más poblados.

Como se ha mencionado anteriormente, los valores de probabilidad del agua subterránea de exceder el umbral de CE de 800 µS/cm son más elevados al noroeste y al sureste de la zona de estudio y, por tanto, las aguas subterráneas de dicha zona tienen más probabilidad de presentar una salinidad elevada no apta (o recomendada) para el consumo humano. En el noroeste se encuentran varios núcleos urbanos que se podrían ver afectados por valores altos de salinidad de las aguas subterráneas como, de oeste a este, Kayes, Nioro, Niono, Mopti y Douentza, así como varias zonas en las que no se supera los 5 habitantes/ha de densidad. En el sureste, el único núcleo urbano que presenta una probabilidad elevada de verse afectada por la salinidad de las aguas subterráneas es el municipio de Anefif. Por el contrario, en la zona suroccidental, donde se ubican los núcleos urbanos más importantes, la probabilidad de exceder el umbral de CE de 800 µS/cm es muy baja.

5.CONCLUSIONES

En este trabajo se han estimado, mediante cartografías predictivas, las regiones de la República de Mali que podrían verse afectadas por una alta salinidad de las aguas subterráneas. Los resultados obtenidos mediante algoritmos de inteligencia artificial muestran que la salinidad de las aguas subterráneas es elevada en zonas con temperaturas elevadas y bajas precipitaciones. Esto puede deberse a que las altas temperaturas incrementan la tasa de evapotranspiración y las bajas precipitaciones derivan en una baja recarga de agua subterránea y, por tanto, en último término desemboca en un incremento de la salinidad de las aguas subterráneas. Además, los resultados muestran una clara relación inversa entre la elevación del terreno y la salinidad de las aguas subterráneas.

Los algoritmos “basados en árboles” utilizados para realizar las cartografías predictivas muestran que las variables explicativas más importantes a la hora de condicionar la conductividad eléctrica de las aguas subterráneas son principalmente la temperatura, la recarga, la hidrogeología, la evapotranspiración real, la precipitación, la elevación y el tipo de suelo. El empleo de diferentes umbrales discriminativos entre clases para la conductividad eléctrica ha mostrado que los resultados óptimos se obtienen con los valores de 500 y 800 µS/cm, descartando así los de 1500 y 2500 µS/cm. Para el umbral de CE de 500 µS/cm, los algoritmos que mejor rendimiento han obtenido para predecir la conductividad eléctrica son ETC, RFC, GBC y LDA, los cuales pertenecen a la familia de algoritmos “basados en árboles”, excepto LDA. Estos algoritmos han sido capaces de distinguir entre la clase mayoritaria (0) y la clase minoritaria (1) sin necesidad de emplear una técnica de remuestreo. Para el caso del umbral de CE de 800 µS/cm, los algoritmos que mejor rendimiento han mostrado son GBC, LDA y RFC junto con la técnica de remuestreo SMOTE. Para los umbrales de CE de 1500 y 2500 µS/cm, los algoritmos no han sido capaces de distinguir entre la clase mayoritaria y la clase minoritaria aún en los casos donde se ha empleado un sobremuestreo de la clase minoritaria, por lo que los resultados para estos umbrales de CE no han sido empleados para elaborar las cartografías.

Las regiones que presentan una mayor probabilidad de exceder los umbrales de conductividad eléctrica se localizan en el sector noroccidental y suroriental de la zona de estudio. Los núcleos urbanos más importantes se sitúan al suroeste del país y, en cambio, las zonas del noroeste y sureste del área de estudio se encuentran relativamente menos pobladas. Por tanto, los núcleos urbanos importantes potencialmente afectados por el consumo de agua subterránea con una salinidad por encima de 800 µS/cm son Kayes, Nioro, Niono, Mopti y Douentza al noroeste, y Anefif al sureste.

Las cartografías obtenidas como resultado final de este trabajo pueden ser de gran utilidad en las tareas de planificación y gestión de los recursos hídricos a escala regional. Este tipo de enfoques, basados en algoritmos de inteligencia artificial, abren una nueva dimensión a la hora de acometer problemáticas relacionadas con el agua en zonas en las que se disponga de la información necesaria.

6. REFERENCIAS BIBLIOGRÁFICAS

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Scientific data, 5 (1), 170191. https://doi.org/10.1038/sdata.2017.191.

Akramkhanov, A., & Vlek, P.L. (2012). The assessment of spatial distribution of soil salinity risk using neural network. Environ. Monit. Assess., 184, 2475–2485.

Akter, F., Bishop, T.F.A., Vervoort, R.W. (2021). Space-time modelling of groundwater level and salinity. Science of the Total Environment, 776. https://doi.org/10.1016/j.scitotenv.2021.145865.

Alagha, J.S., Seyam, M., Said, M.A.M., Mogheir, Y. (2017). Integrating an artificial intelligence approach with k-means clustering to model groundwater salinity: the case of Gaza coastal aquifer (Palestine). Hydrogeol. J., 25, 2347–2361.

Araya, D., Podgorski, J., Berg, M. (2023). Groundwater salinity in the Horn of Africa: Spatial prediction modeling and estimated people at risk. Environment International, 176, 107925, ISSN 0160-4120.

Attia, A., Qureshi, A.S., Kane, A.M., Alikhanov, B., Kheir, A.M.S., Ullah, H., Datta, A., Samasse, K. (2022). Selection of Potential Sites for Promoting Small-Scale Irrigation across Mali Using Remote Sensing and GIS. Sustainability, 14, 12040. https://doi.org/10.3390/su141912040.

Banerjee, P., Singh, V.S., Chatttopadhyay, K., Chandra, P.C., Singh, B. (2011). Artificial neural network model as a potential alternative for groundwater salinity forecasting. J. Hydrol, 398, 212–220.

Barzegar, R., & Moghaddam, A.A. (2016). Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Modeling Earth Syst. Environ. 2, 26. https://doi.org/10.1007/s40808-015-0072-8.

Bourke, S.A., Hermann, K.J., Hendry, M.J. (2017). High-resolution vertical profiles of groundwater electrical conductivity (EC) and chloride from direct-push EC logs. Hydrogeol J, 25: 2151–2162.

British Geological Survey (BGS). Africa Groundwater Atlas, Geology, disponible en: http://earthwise.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res., 16, 321–357. https://doi.org/10.1613/jair.953.

Cuthbert, M.O., Taylor, R.G., Favreau, G., Todd, M.C., Shamsudduha, M., Villholth, K.G., MacDonald, A.M., Scanlon, B.R., Kotchoni, D.O.V., Vouillamoz, J.-M., Lawson, F.M.A., Adjomayi, P.A., Kashaigili, J., Seddon, D., Sorensen, J.P.R., Ebrahim, G.Y., Owor, M., Nyenje, P.M., Nazoumou, Y., Goni, I., Ousmane, B.I., Sibanda, T., Ascott, M.J., Macdonald, D.M.J., Agyekum, W., Koussoubé, Y., Wanke, H., Kim, H., Wada, Y., Lo, M.-H., Oki, T., Kukuric, N., (2019). Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. Nature, 572, 230–234. https://doi.org/10.1038/s41586-019-1441-7.

Delsman, J.R,, Van Baaren, E.S., Siemon, B., Dabekaussen, W., Karaoulis, M.C., Pauw, P.S., Vermaas, T., Bootsma, H., De Louw, P.G.B., Gunnink, J.L., Wim Dubelaar, C., Menkovic, A., Steuer, A., Meyer, U., Revil, A., Oude Essink, G.H.P. (2018). Large-scale, probabilistic salinity

mapping using airborne electromagnetics for groundwater management in Zeeland, the Netherlands. Environ Res Lett, 13. https://doi.org/10.1088/1748-9326/aad19e

Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Michaeli, E., Montanarella, L., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmore, R. (2013). Harmonisation of the soil map of Africa at the continental scale. Geoderma, 211-212, 138-153. https://doi.org/10.1016/j.geoderma.2013.07.007, 2013.

Díaz-Alcaide, S., Martínez-Santos, P., Villarroya, F. (2017). A Commune-Level Groundwater Potential Map for the Republic of Mali. Water, 9, 839, doi:10.3390/w9110839.

Direction Nationale de l'Hydraulique. (2010). Données Hydrogeologiques et des Forages. Direction Nationale de l'Hydraulique.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36, 27-46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

European Space Agency (ESA). (2010). GlobCover 2009 (Global Land Cover Map), disponible en: http://due.esrin.esa.int/page_globcover.php (último acceso: 28 de Junio de 2024).

Foster, S., Tuinhof, A., Garduño, H. (2006). Sustainable Groundwater Management. Lessons from Practice, Case profile collection Groundwater Development in Sub-saharan Africa. A Strategic Overview of Key Issues and Major Needs, vol. 15, Groundwater Management Advisory Team (GW-MATE), World Bank.

Geng, X. & Boufadel, M.C. (2017). The influence of evaporation and rainfall on supratidal groundwater dynamics and salinity structure in a sandy beach. Water Resource Research, 53, 6218-6238.

George, R.J., McFarlane, D.J., Nulsen, R.A. (1997). Salinity threatens the viability of agriculture and ecosystems in Western Australia. Hydrogeology Journal, 5, 6-21.

Géron, A. (2019). Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow, 2ª edición. Ed. Anaya.

Gholami, V., Khaleghi, M.R., Sebghati, M. (2017). A method of groundwater quality assessment based on fuzzy network-CANFIS and geographic information system (GIS). Appl. Water Sci., 7(7), 3633-3647.

Gil-Márquez, J.M., Barberá, J.A., Andreo, B., Mudarra, M. (2017). Hydrological and geochemical processes constraining groundwater salinity in wetland areas related to evaporitic (karst) systems. A case study from southern Spain. J Hydrol, 544, 538–554.

Gleeson, T., Wada, Y., Bierkens, M.F., Van Beek, L.P. (2012). Water balance of global aquifers revealed by groundwater footprint. Nature, 488, 197-200.

Gómez-Escalonilla, V., Martínez-Santos, P., Martín-Loeches, M. (2022). Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali. Hydrology and Earth System Sciences, 26, 221-243. https://doi.org/10.5194/hess-26-221-2022.

Gómez-Escalonilla, V. (2024). Metodologías de aprendizaje automático para la optimización de campañas de prospección hidrogeológica y mejora del acceso al agua en el Sahel. Tesis Doctoral. Universidad Complutense de Madrid. 376 pp.

Golchin, I., & Azhdary Moghaddam, M. (2016). Hydro-geochemical characteristics and groundwater quality assessment in Iranshahr plain aquifer, Iran. Environmental Earth Sciences, 75, 1-14.

Haggerty, R., Sun, J., Yu, H., Li, Y. (2023). Application of machine learning in groundwater quality modeling-A comprehensive review. Water Research, 233, 119745.

He, H., Bai, Y., Garcia, E.A., Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969.

Huang, W. & Foo, S. (2002). Neural network modeling of salinity variation in Apalachicola River. Water Res., 36, 356-362.

Hosmer, D.W. & Lemeshow, S. (2000). Area under the ROC curve. Applied logistic regression. 160-164.

Hossin, M. & Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluation. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5 (2), 1-11. DOI : 10.5121/ijdkp.2015.5201.

Huntington, J., Hegewisch, K., Daudert, B., Morton, C., Abatzoglou, J., McEvoy, D., Erickson, T. (2017). Climate Engine: Cloud Computing of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. Bulletin of the American Meteorological Society, 2397-2410. https://doi.org/10.1175/BAMS-D-15-00324.1

Imperato, P. J., Clark A., Baker, K.M. (2024). Mali. Encyclopedia Britannica. Disponible en: https://www.britannica.com/place/Mali (último acceso: 25 de junio 2024).

Iris Rodríguez, C., Duque, C., Calvache, M.L., López-Chicano, M. (2010). Causas de las variaciones de la conductividad eléctrica del agua subterránea en el acuífero Motril-Salobreña, España. Geogaceta, 49, 107-110. ISSN: 2173-6545.

Jones, P.D. & Harris, I. (2013). CRU TS3.21: Climatic Research Unit (CRU) Time-Series (TS) Version 3.21 of High resolution gridded data of month-by-month variation in climate (Jan. 1901-Dec. 2012). NCAS British Atmospheric Data Centre, 24th September 2013. https://dx.doi.org/10.5285/D0E1585D-3417-485F-87AE-

Lemaître, G., Nogueira, F., Aridas, C.K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J. Mach. Learn. Res., 18, 1-5.

Levanon, E., Yechieli, Y., Gvirtzman, H., Shalev, E. (2017). Tide-induced fluctuations of salinity and groundwater level in unconfined aquifers, field measurements and numerical model. J Hydrol., 551, 665-675.

MacDonald, A.M., Bonsor, H.C., Dochartaigh, B.É.Ó., Taylor, R.G. (2012). Quantitative maps of groundwater resources in Africa. Environmental Research Letters, 7, 024009. 7 pp. https://doi.org/10.1088/1748-9326/7/2/024009

MacDonald, A.M., Lark, R.M., Taylor, R.G., Abiye, T., Fallas, H.C., Favreau, G., Goni, I.B., Kebede, S., Scanlon, B., Sorensen, J.P.R., Tijani, M., Upton, K.A., West, C. (2021). Mapping groundwater recharge in Africa from ground observations and implications for water security. Environmental Research Letters, 16 (3), 034012. https://doi.org/10.1088/1748-9326/abd661.

Masciopinto, C., Liso, I.S., Caputo, M.C., De Carlo, L. (2017). An integrated approach based on numerical modelling and geophysical survey to map groundwater salinity in fractured coastal aquifers, Water, 9 (11), 875, https://doi.org/10.3390/w9110875.

Mosavi, A., Hosseini, F. S., Choubin, B., Goodarzi, M., & Dineva, A. A. (2020). Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. IEEE Access, 8, 145564-145576. DOI:10.1109/ACCESS.2020.3014908

Mosavi, A., Sajedi Hosseini, F., Choubin, B., Taromideh, F., Ghodsi, M., Nazari, B., Dineva, A.A. (2021). Susceptibility mapping of groundwater salinity using machine learning models. Environ. Sci. Pollut. Res. 28, 10804-10817. https://doi.org/ 10.1007/s11356-020-11319-5.

Motevalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V. (2019). Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. Journal of Cleaner Production, 228, 1248-1263.

Muthusi, F., Mahamud, G., Abdalle, A., Gadain, H. (2007). Rural Water Supply Assessment. Technical Report No-08, FAO-SWALIM, Nairobi, Kenya.

NASA Shuttle Radar Topography Mission (SRTM). (2013). Shuttle Radar Topography Mission (SRTM) Global. Disponivle en: OpenTopography. https://doi.org/10.5069/G9445JDF (último acceso: 25 de junio de 2024).

Nosetto, M.D., Acosta, A.M., Jayawickreme, D.H., Ballesteros, S.I., Jackson, R.B., Jobbágy, E.G. (2013). Land-use and topography shape soil and groundwater salinity in central Argentina. Agricultural Water Management, 129: 120-129.

Oficina de Información Diplomática del Ministerio de Asuntos Exteriores: Ficha país, República de Mali. (2023). Disponible en: https://www.exteriores.gob.es/Documents/FichasPais/MALI_FICHA%20PAIS.pdf (último acceso: 18 de junio 2024).

Pauw, P.S., Groen, J., Groen, M.M.A., van der Made, K.J., Stuyfzand, P.J., Post, V.E.A. (2017). Groundwater salinity patterns along the coast of the Western Netherlands and the application of cone penetration tests. Journal of Hydrology, 551: 756-767. https://doi.org/10.1016/j.jhydrol.2017.04.021

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Peel, M.C., Finlayson, B.L., McMahon, T.A. (2007). Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth System Sciences, 11, 1633-1644. https://doi.org/10.5194/hess-11-1633-2007

Poggio, L., de Sousa, L. M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL, 7, 217-240. https://doi.org/10.5194/soil-7-217-2021.

Rasse, M. (2010). Carte Geologique du Mali. Atlas Mali Jeune Afrique. Disponible en: https://www.researchgate.net/publication/258555891_Carte_Geologique_du_Mali (último acceso: 21 de junio 2024).

Sahour, H., Gholami, V., Vazifedan, M. (2020). A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. Journal of Hydrology, 591, 125321. ISSN: 0022-1694, https://doi.org/10.1016/j.jhydrol.2020.125321

Salama, R.B., Otto, C.J., Fitzpatrick, R.W. (1999). Contributions of groundwater conditions to soil and water salinization. Hydrogeology Journal, 7, 46-64. https://doi.org/10.1007/s100400050179

Scanlon, B.R., Gates, J.B., Reedy, R.C., Jackson, W.A., Bordovsky, J.P. (2010). Effects of irrigated agroecosystems: 2. Quality of soil water and groundwater in the southern High Plains, Texas. Water Resources Research, 46, W09538. doi:10.1029/2009WR008428

Schlüter, T. (2006). Geological Atlas of Africa: With Notes on Stratigraphy, Tectonics, Economic Geology, Geohazards and Geosites of each country, 255 pp, Springer Berlin Heidelberg.

Smedley, P. (2002). Groundwater Quality: Mali. British Geological Survey and Water Aid: 1-5. Disponible en: http://nora.nerc.ac.uk/516317/ (último acceso: 25 de junio 2024).

Suthaharan, S. (2016). Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Integrated Series in Information Systems. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4899-7641-3

Thanh, N. N., Thunyawatcharakul, P., Ngu, N. H., Chotpantarat, S. (2022). Global review of groundwater potential models in the last decade: parameters, model techniques, and validation. Journal of Hydrology, 614 (2), 128501. DOI:10.1016/j.jhydrol.2022.128501

Traore, A. Z., Bokar, H., Sidibe, A., Upton, K., Ó Dochartaigh, B., Bellwood-Howard, I. (2018). Africa Groundwater Atlas: Hydrogeology of Mali, disponible en: http://earthwise.bgs.ac.uk/index.php/ Hydrogeology_of_Mali (último acceso: 17 de junio 2024), 2018.

United Nations, Department of Economic and Social Affairs, Population Division. (2022). World Population Prospects 2022, Online Edition.

United Nations. (2021). The United Nations World Water Development Report 2021: Valuing Water. UNESCO, París.

United Nations. (2002). General comment no. 15 (2002). The right to water (arts. 11 and 12 of the International Covenant on Economic, Social and Cultural Rights). COMMITTEE ON ECONOMIC, SOCIAL AND CULTURAL RIGHTS, Geneva, 11-29 November 2002.

UNICEF/WHO. (2022). Progress on drinking water, sanitation and hygiene in Africa 2000-2020: Five years into the SDGs. United Nations Children's Fund (UNICEF) and World Health Organization (WHO), Nueva York.

UNICEF/WHO. (2021). Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), Ginebra. Licence: CC BY-NC-SA 3.0 IGO.

Wada, Y., Van Beek, L.P., Van Kempen, C.M., Reckman, J.W., Vasak, S., Bierkens, M.F. (2010). Global depletion of groundwater resources. Geophysical Research Letters, 37.

WHO, W.H.O. (2011). Guidelines for Drinking-water Quality. Ginebra, Suiza.

WorldPop. (2017). Mali - Population density (2015), disponible en: https://energydata.info/dataset/mali-republic-population-density-2015/resource/5388ca5d-e70f-4998-966c-95f5264b5178 (último acceso: 24 de julio 2024).

Xie, Y., Sha, Z., Yu, M. (2008). Remote sensing imagery in vegetation mapping: a review. Journal of Plant Ecology, 1, 9-23. https://doi.org/10.1093/jpe/rtm005.

Xu, H. (2006). Modification of Normalized Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery, International Journal of Remote Sensing, 27, 3025-3033. https://doi.org/10.1080/01431160600589179.

Zheng, A. & Casari, A. (2018). Feature Engineering for Machine Learning. O'Reilly Media, Inc. ISBN: 9781491953242