Facultad de Ciencias Geológicas

Universidad Complutense de Madrid

MÁSTER UNIVERSITARIO EN GEOLOGÍA AMBIENTAL

Curso 2024-2025

Cartografía predictiva de la conductividad eléctrica de las aguas subterráneas mediante herramientas de inteligencia artificial en el sur de Madagascar y estimación de personas en riesgo.

Predictive mapping of groundwater electrical conductivity by means of machine learning tools in southern Madagascar and estimation of people at risk.

BÁRBARA LÓPEZ LUNA

TUTOR DEL TRABAJO: VÍCTOR GÓMEZ-ESCALONILLA CANALES

Facultad de Ciencias Geológicas

Universidad Complutense de Madrid

MÁSTER UNIVERSITARIO EN GEOLOGÍA AMBIENTAL

Curso 2024-2025

Cartografía predictiva de la conductividad eléctrica de las aguas subterráneas mediante herramientas de inteligencia artificial en el sur de Madagascar y estimación de personas en riesgo.

Predictive mapping of groundwater electrical conductivity by means of machine learning tools in southern Madagascar and estimation of people at risk.

BÁRBARA LÓPEZ LUNA

TUTOR DEL TRABAJO: VÍCTOR GÓMEZ-ESCALONILLA CANALES

Fdo.:

GOMEZ-ESCALONILLA CANALES VICTOR - 48283243X

Firmado digitalmente por GOMEZ-ESCALONILLA CANALES VICTOR - 48283243X
Fecha: 2025.07.28 00:30:24 +02'00'

DECLARACION DE NO PLAGIO

D./Dña. BÁRBARA LÓPEZ LUNA con NIF: 05449934S
Estudiante de Máster Universitario en Geología Ambiental de la Facultad de Geología de la Universidad Complutense de Madrid, curso 2024 /2025 como autor/a de este documento académico titulado: “Cartografía predictiva de la conductividad eléctrica de las aguas subterráneas mediante herramientas de inteligencia artificial en el sur de Madagascar y estimación de personas en riesgo” y presentado como Trabajo Fin de Máster, para la obtención del título correspondiente, cuyo tutor es VÍCTOR GÓMEZ-ESCALONILLA CANALES.

DECLARO QUE:
El Trabajo de Fin de Máster que presento está elaborado por mí, 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. Asimismo, no he hecho uso de información no autorizada de cualquier fuente escrita, de otra persona, de trabajo escrito de otro o cualquier otra fuente.
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 2025

Fdo.: 

UNIVERSIDAD COMPLUTENSE MADRID

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

Yo, LÓPEZ LUNA, BÁRBARA

Con DNI: 05449934S

Declaro de manera responsable que el/la presente:

- Trabajo de Fin de Máster (TFM)

Titulado/a

"CARTOGRAFÍA PREDICTIVA DE LA CONDUCTIVIDAD ELÉCTRICA DE LAS AGUAS SUBTERRÁNEAS MEDIANTE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL EN EL SUR DE MADAGASCAR Y ESTIMACIÓN DE PERSONAS EN RIESGO"

Es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo con los 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 2025

Fdo.: 

Agradecimientos

Mi primer agradecimiento se lo dedico a mi tutor de TFM, Víctor Gomez-Escalonilla, por brindarme su tiempo hasta el punto de tener tutorías semanales para ir revisando mi trabajo, por tener la paciencia de orientarme cuando no sabía muy bien como continuar, por su amabilidad y comprensión, así como su ayuda cuando las cosas se torcían. Por qué, aún sin tener la obligación de comportarse de ese modo, lo hizo, y no todo el mundo lo hace. Por eso se lo agradezco, por que si no hubiera sido por su apoyo seguramente este trabajo no hubiera salido adelante, al menos no este año. Se merece un aumento de sueldo.

A aquellos profesores a lo largo de mi carrera universitaria que me inspiraron a seguir adelante, siendo uno de ellos Pedro Martínez Santos, cuyas clases de hidrogeología me condujeron hasta este master, iniciando el camino a mi futura vocación.

A mi familia, la cual me ha estado apoyando todos estos años, estando siempre ahí para mí, en los buenos y malos momentos, brindándome su apoyo incondicional, sacrificando muchas veces su tiempo y energía para que pudiera alcanzar mis metas, por creer en mi cuando a veces ni si quiera yo misma creía en mí. Os quiero un montón.

A mi mejor amigo Junze Chen, con el que he tenido el placer de compartir tanto carrera como master, por estar acompañándome 6 años en esta ardua experiencia, echándome una mano cada vez que lo necesitaba, por las risas compartidas y los momentos inolvidables, por dejarte llevar y acompañarme en mis locuras hasta el punto de viajar a la otra punta del mundo. Gracias, soy muy afortunada de contar con tú amistad.

A Mari, la cual de un año para otro se convirtió en un gran apoyo para mí, ayudándome en mis momentos más difíciles, animándome y sacándome siempre una sonrisa.

A los míos, a todos los amigos que han estado para mi estos años, amigos de carrera, amigos del barrio y amigos del master. Gracias por estar ahí, aunque fuera para ir a comer una hamburguesa, ir al gimnasio o simplemente dar un paseo. Aprecio mucho el tiempo que he pasado con vosotros. Vuestra compañía me hacía disfrutar de mi día a día y me ayudaba a seguir a delante.

A mis compis del trabajo, por hacerme los turnos más amenos y siempre hacerme el favor de cambiarme los horarios si necesitaba algún día para mi master.

Por último, a una persona muy importante y que muchas veces se me olvida, a mí misma, por que si no hubiera seguido adelante no estaría aquí ahora mismo.

ÍNDICE

RESUMEN

El sur de la República de Madagascar se caracteriza por un clima tropical seco, con una prolongada estación seca y cálidas temperaturas. En este contexto las aguas subterráneas suponen un sustento fundamental para la población, sobre todo en aquellas regiones situadas en la costa suroccidental, que registra las condiciones más áridas. A su vez, a las particularidades de tratarse de una isla, rodeada por agua salada, pueden provocar la intrusión de la cuña salina. Esto puede afectar a las aguas subterráneas, por ello, dadas las circunstancias, uno de los principales problemas que pueden presentar las aguas subterráneas en esta región es un exceso de salinidad, pudiendo limitar su aptitud para el consumo humano. Por consiguiente, en este proyecto se ha elaborado una cartografía predictiva, mediante la utilización de algoritmos de inteligencia artificial, de la conductividad eléctrica de las aguas subterráneas en la región meridional de Madagascar. La cartografía final obtenida como resultado, puede ser de gran utilidad proporcionando información muy relevante a los organismos competentes para tratar de mejorar el acceso al agua potable para la población. Para ello, se ha empleado una base de datos conformada por más de 2300 pozos de agua con datos de conductividad eléctrica. Además, se ha generado una base de datos de 20 variables explicativas que pueden condicionar la conductividad eléctrica del agua subterránea, incluyendo factores climáticos, geológicos, edafológicos y topográficos, entre otros. Se han evaluado tres umbrales diferentes de conductividad eléctrica para diferenciar entre puntos aptos y no aptos, siendo el umbral intermedio, de 1000 µS/cm, el que mejores resultados generales obtuvo. Los resultados indican que las zonas situadas en la costa oeste y en el sur del área de estudio son aquellas con mayor probabilidad de tener una conductividad eléctrica elevada. A su vez, se han podido distinguir variables explicativas con un alto peso a la hora de influir a la conductividad eléctrica, como la precipitación, evapotranspiración real media y la evapotranspiración real en los periodos húmedos, junto con variables de distancia y topografía, en concreto distancia a la costa y elevación del terreno. Por último, se ha realizado un análisis de la población y de las principales aldeas situadas en zonas donde la cartografía indica una alta conductividad eléctrica, resultando en que el 33% de la población se localiza en zonas con una probabilidad superior al 50% de que la conductividad eléctrica supere el umbral de 1000 µS/cm.

1.INTRODUCCIÓN

El agua potable es una necesidad humana básica, imprescindible para beber, el saneamiento, la higiene y la seguridad alimentaria (UNESCO, 2021). Se trata de un derecho humano fundamental para el desarrollo de una vida digna (United Nations, 2002). Sin embargo, hoy en día, 884 millones de personas en todo el mundo carecen aún tanto del acceso al propio recurso como a aguas de buena calidad. Esto afecta mayoritariamente a comunidades menos desarrolladas. Además, factores como el cambio climático y el crecimiento demográfico aumentan la demanda de los recursos hídricos lo que puede aumentar la dificultad de acceso al agua potable en un futuro, provocando un "estrés hídrico" aún más alto (ONU-Hábitat, s. f.). En África, el continente con mayores dificultades para el abastecimiento de agua potable, únicamente un 39% de la población disponía del acceso a fuentes mejoradas de agua potable y libre de posible contaminación en el año 2020. Por ello, es imprescindible implementar medidas que permitan obtener fuentes seguras de agua potable cercanas a las poblaciones para mejorar el acceso a este recurso vital. Actualmente, Madagascar es uno de los países dentro del continente africano más afectados por la crisis del agua. Los últimos datos indican que un 66% de la población de zonas rurales carece de un servicio básico de agua potable frente al 49% de la población urbana (Unicef, 2018; UNICEF/WHO, 2022). Por lo general, las regiones del sur de Madagascar, zona en la que se desarrolla este trabajo, tienen una de las coberturas más bajas de acceso al agua potable. Se trata de una región muy vulnerable a las sequías, dando lugar a llamamientos de emergencia anuales para salvar vidas de niños con altos niveles de desnutrición (Serele et al., 2020). Además, existe una notable desigualdad entre el acceso al agua potable en función del nivel de riqueza. Únicamente el 20% de las poblaciones más pobres disponen de acceso a un servicio básico de agua potable frente al 83% atendiendo a las poblaciones más ricas (Sanitation and Water for All, 2022).

En regiones áridas y semiáridas, las aguas subterráneas suponen un recurso vital de abastecimiento. Se trata de la principal fuente de abastecimiento para miles de millones de personas alrededor de todo el mundo, desempeñando un papel fundamental tanto en la agricultura de regadío, como en la salud humana y en el mantenimiento de otros ecosistemas dependientes de este recurso (Sahuquillo, 2009; UNESCO, 2022). Sin embargo, en la actualidad las aguas subterráneas presentan globalmente un agotamiento insostenible debido a la sobreexplotación del ser humano (Gleeson et al., 2012). En las zonas áridas y semiáridas, con escasez de recursos hídricos superficiales, como en la zona de estudio, el agua subterránea juega un papel fundamental, convirtiéndose en muchos casos en la única fuente confiable de agua dulce, especialmente en los periodos más secos. En la región meridional de Madagascar, el clima árido junto con las precipitaciones irregulares afecta a las

comunidades menos resilientes, en donde la economía depende principalmente de la agricultura, lo que aumenta la dependencia de la disponibilidad de las aguas subterráneas en cantidad y calidad suficientes (Serele et al., 2020). Sin embargo, la sobreexplotación de este recurso puede provocar cambios erráticos en la calidad de agua subterránea, sobre todo en acuíferos costeros, ya que se pueden dar procesos de intrusión salina, aumentando la conductividad eléctrica y empeorando la calidad del recurso (Sahour et al., 2020).

Por ello, en este contexto, la elaboración de cartografías predictivas para localizar zonas con valores de salinidad aptos para el consumo humano pueden ser una gran herramienta para mejorar el acceso a agua potable de buena calidad. En las últimas décadas, este tipo de cartografías predictivas se han desarrollado a través de técnicas de aprendizaje automático (*Machine Learning*). Según Sahour (2020), este método automatiza un marco de modelos analíticos. El aprendizaje automático es una rama perteneciente a la inteligencia artificial que se basa en el concepto de que los sistemas pueden aprender de datos, reconocer patrones y tomar decisiones con la mínima intervención humana (BBVA, 2024; Sahour et al., 2020). La aplicación de algoritmos y modelos de inteligencia artificial se ha visto incrementado en los últimos años en el campo de la hidrogeología (Rajaee et al., 2019; Adombi et al., 2021). Estos estudios se han centrado principalmente en la elaboración de cartografías de potencial hidrogeológico (Gómez-Escalonilla et al., 2022), predicción de niveles piezométricos (Martínez-Santos et al., 2025), predicción de contaminantes en las aguas subterráneas (Araya et al., 2023; Gómez-Escalonilla et al., 2024) y la cartografía de ecosistemas dependientes de aguas subterráneas (Martínez-Santos et al., 2021). En relación con el campo de este trabajo, los estudios previos que emplearon inteligencia artificial para modelar la salinidad en las aguas subterráneas se enfocaron principalmente en el uso de redes neuronales artificiales (RNA) (Huang y Foo, 2002; Akramkhanov y Vlek, 2012; Alagha et al., 2017; Barzegar y Moghaddam, 2016; Jeong et al., 2024). Estas técnicas han mostrado eficacia en la elaboración de mapas de salinidad. No obstante, su carácter de “caja negra” dificulta la interpretación y cuantificación del impacto de los distintos factores de control involucrados en el proceso de modelado (Malaxetxebarria Bengoetxea, 2024; Sahour et al., 2020).

Por otro lado, diversos estudios han empleado modelos pertenecientes a la familia de algoritmos denominados “basados en árboles”, entre los cuales destacan técnicas como Random Forest (Akter et al., 2021; Mosavi et al., 2021) y Gradient Boosting (Sahour et al., 2020). Estos algoritmos son especialmente valorados por su capacidad para manejar datos complejos y relaciones no lineales entre variables. Gracias a su estructura, estos métodos no solo ofrecen predicciones precisas, sino que también permiten evaluar y cuantificar la relevancia o importancia de cada variable dentro del modelo. Esta característica facilita una mejor interpretación del comportamiento interno del modelo, brindando a los investigadores

una comprensión más profunda de los factores que influyen en los resultados, tal como señalan Gómez-Escalonilla (2024) y Malaxetxebarria Bengoetxea (2024).

1.1Objetivos

Este trabajo presenta varios objetivos a considerar. El primero es la elaboración de una cartografía predictiva de la conductividad eléctrica de las aguas subterráneas de la zona sur de la República de Madagascar. Para ello, será necesario integrar las bases de datos de puntos de agua existentes junto a posibles factores que afecten a este parámetro en un Sistema de Información Geográfica. Posteriormente, se aplicarán técnicas y algoritmos de *machine learning* (ML) que permitan obtener una buena capacidad predictiva y que cuyos resultados concuerden con los modelos conceptuales preexistentes en la zona o en otras zonas con factores similares.

Por último, el objetivo final de este trabajo consistirá en realizar un análisis que permita estimar la población en riesgo de consumir aguas con altas salinidades, con el fin de señalar las regiones habitadas más vulnerables o en riesgo de superar cierto umbral de conductividad eléctrica.

2.CONTEXTUALIZACIÓN GENERAL DE LA ZONA DE ESTUDIO

2.1 Contexto geográfico y climático

La República de Madagascar (Figura 1) es un país insular perteneciente a África situado en el océano Índico suroccidental, encontrándose a 300 km al este de la costa africana, en la zona intertropical (Upton, 2018). Abarca una extensión total de 587.041 km² siendo la quinta isla más grande del mundo. Presenta una población total de 31.195.932, de los cuales el 40.6% vive en zonas urbanas (World Bank, 2025), siendo aquella con mayor número población la capital, Antananarivo, con 3.872.000 habitantes. Presenta una densidad poblacional de 47.7 habitantes por kilómetro cuadrado (Oficina de Información Diplomática del Ministerio de Asuntos Exteriores, 2025).

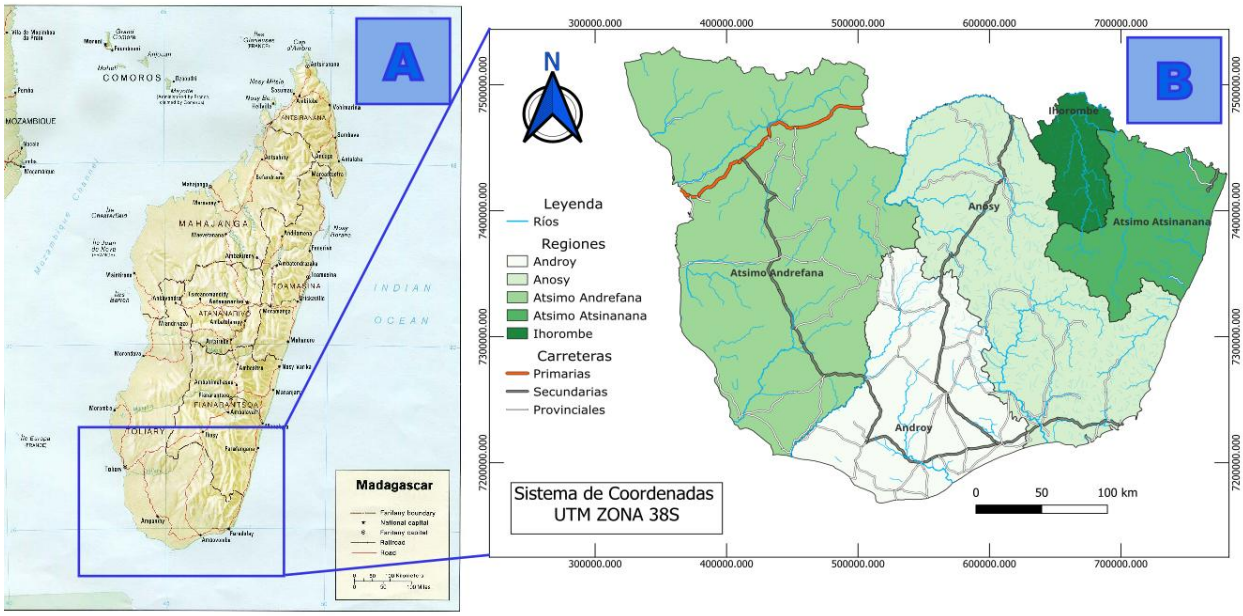

Figura 1: A) Mapa topográfico de Madagascar (Modificado a partir de United States Central Intelligence Agency, 1981). B) Mapa de las regiones de Madagascar pertenecientes a la zona de estudio con sus principales elementos geográficos y redes de drenaje.

Esta se encuentra formada por 6 provincias y 23 regiones administrativas, a su vez las regiones se encuentran divididas en distritos, comunas y *fokontany* (barrios o aldeas).
Geográficamente se divide en dos regiones diferentes geomorfológicamente. Presenta una zona montañosa la cual abarca el 67% de la isla, la cual se alza bruscamente en la zona costera del este, disminuyendo hacia el oeste. Esta zona está formada por rocas del basamento Precámbrico, presentando valles fluviales y colinas redondeadas. La altitud media de la zona es de 2000 m, siendo el macizo de Tsaratanana el más elevado topográficamente, alcanzando los 2876 m en el monte Maromokotro.

A su vez presenta tierras bajas en los alrededores costeros. Estas cuencas sedimentarias cubren el 33% restante de la isla (Oficina de Información Diplomática del Ministerio de Asuntos Exteriores, 2025; Upton, 2018).

Entre los cuerpos de agua superficial pertenecientes a la zona de estudio caben destacar los 5 ríos más importantes para estas regiones, el río Onilahy, Mangoky, Fiherenana, Linta y Menarandra, siendo los ríos Mangoky y Onilahy unos de los más largos en la isla (WorldAtlas, 2025).

El clima de Madagascar es variable (Figura 2), presentando un clima tropical en la costa este, templado en la meseta interior y árido en el extremo sur. Presenta una precipitación anual media de 1700 mm (Figura3), pero varía dependiendo de la región, registrándose la máxima pluviometría, cercana a 3000 mm/año en el este y la mínima alrededor de los 400 mm/año en el extremo sur. Por otro lado, la precipitación también presenta una variabilidad intraanual condicionada por la temporada de lluvias asociadas a los monzones del noroeste. Esta abarca desde el mes de noviembre a abril, mientras que la temporada seca, controlada por los cálidos vientos alisios del sureste, abarca desde el mes de mayo hasta octubre. La temperatura media anual es de 17.8ºC, aunque se registra un amplio abanico térmico con temperaturas más cálidas en las costas occidentales y más tenues en las zonas montañosas (Smedley, 2002).

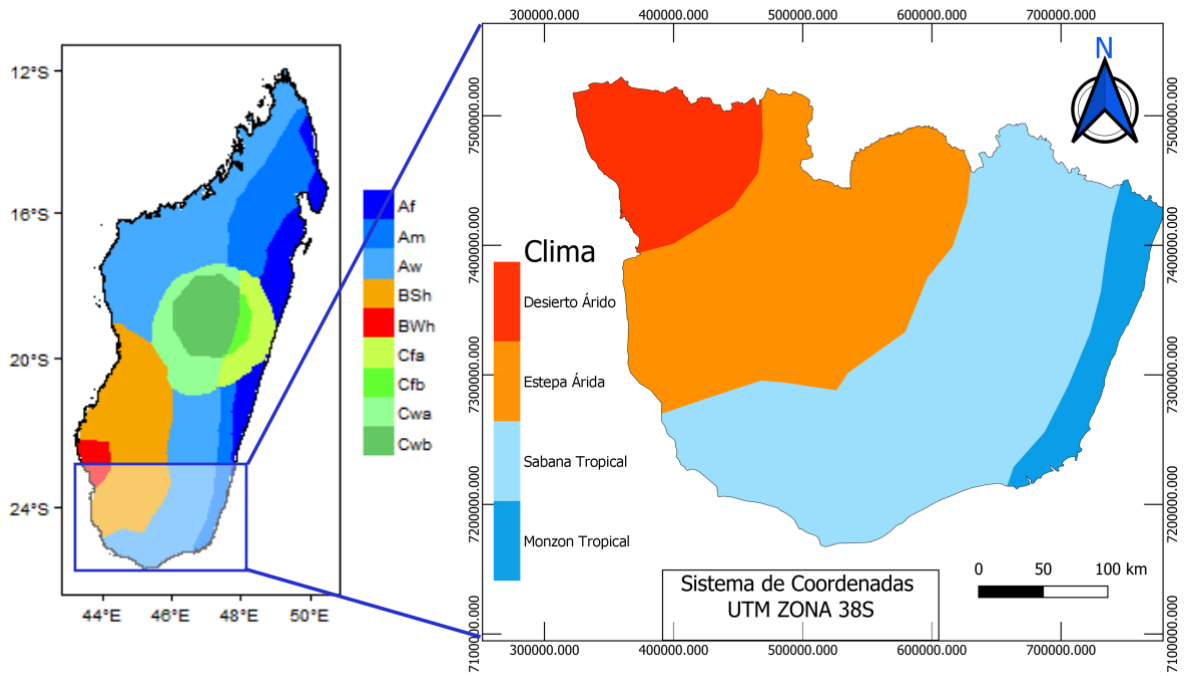

Figura 2: Mapa de zonas climáticas de la zona de estudio según la clasificación de Köppen Geiger (Modificado de Jones & Harris, 2013).

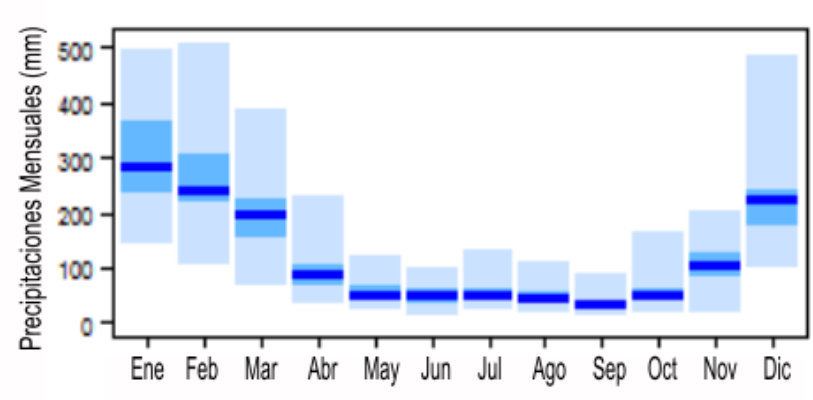

Figura 3: Precipitaciones medias mensuales de Madagascar en mm. Siendo el color azul oscuro la media, el azul medio los percentiles de 25 y 75 y el color azul clarito de precipitaciones máximas y mínimas (modificado de Jones & Harris, 2013).

2.2 Marco geológico e hidrogeológico

La geología de Madagascar tal y como se podrá ver más adelante se compone, en gran medida, por rocas ígneas y metamórficas del basamento de edad precámbrica. Estos materiales, conocidos como “zócalo”, se encuentran formados principalmente por granitos, gneises y esquistos, desde un punto de vista mineralógico, son ricos en grafito y carbón, destacando yacimientos de cobre en la zona suroeste. También se registran formaciones sedimentarias más modernas, conformadas por las cuencas sedimentarias situadas en las zonas más deprimidas y en las regiones costeras, así como aluviones más recientes situados en las zonas montañosas. A su vez, estos depósitos pueden presentar intercalaciones de rocas volcánicas mayoritariamente basálticas. Las cuencas sedimentarias se componen principalmente por secuencias mixtas de arenas, arcillas y materiales carbonatados. La zona occidental presenta sedimentos continentales (Carbonífero Superior – Jurásico), formados principalmente por areniscas, arcillas y conglomerados. Por otro lado, la franja oriental costera presenta abundantes areniscas con algunos depósitos volcánicos (Cretácico – Cuaternario). Por último, en las zonas planas del extremo sur predominan las arcillas y areniscas (Cenozoico), resaltando entre las formaciones sedimentarias más recientes los sistemas dunares del Cuaternario (Smedley, 2002; Upton, 2018).

Desde el punto de vista de la hidrogeología (Figura 4),la zona de estudio presenta distintas tipologías de acuíferos. En la costa oeste y sur se localizan acuíferos no consolidados, formados por aluviones, dunas de arena y arena de playa de productividad variable. Esta varía entre rendimientos bajos-moderados a muy altos, dependiendo de la consolidación y cementación de los sedimentos, presentando espesores de 25 a 30 metros.

En las fronteras entre las regiones de Atsimo-Andrefana y Androy (zona central oeste), se encuentran acuíferos sedimentarios fracturados, caracterizados por una productividad alta, y conformados principalmente por calizas y areniscas que alcanzan espesores de hasta 500 metros. La alta productividad está relacionada con la alta permeabilidad que se da a través de las fracturas del acuífero, lo que da lugar a transmisividades entre 725 m²/d y 6000 m²/d dependiendo de la zona.

Por otro lado, los acuíferos de la región central y oriental, situados en las zonas más elevadas, están compuestos, predominantemente, por materiales del basamento cristalino. Estas rocas forman acuíferos locales de baja productividad, ya que, al estar compuestas por rocas metamórficas e ígneas, a priori de carácter impermeable, su productividad se verá condicionada por la densidad de fracturas, aumentando su capacidad de almacenamiento contra más fracturado se encuentre el material. El espesor del acuífero meteorizado suele ser inferior a 20 metros y presenta un comportamiento libre.

Por último, los acuíferos volcánicos se encuentran puntualmente repartidos por los afloramientos de este tipo de roca que se encuentran en la isla. En la zona de estudio los acuíferos volcánicos muestran una productividad baja a moderada. Apenas existe información relevante sobre estos acuíferos debido a su ubicación geográfica, situándose en zonas remotas o relativamente despobladas. Sin embargo, en aquellas zonas donde se han realizado perforaciones, se ha podido observar que no presentan fracturas, haciéndolas prácticamente impermeables (Upton, 2018).

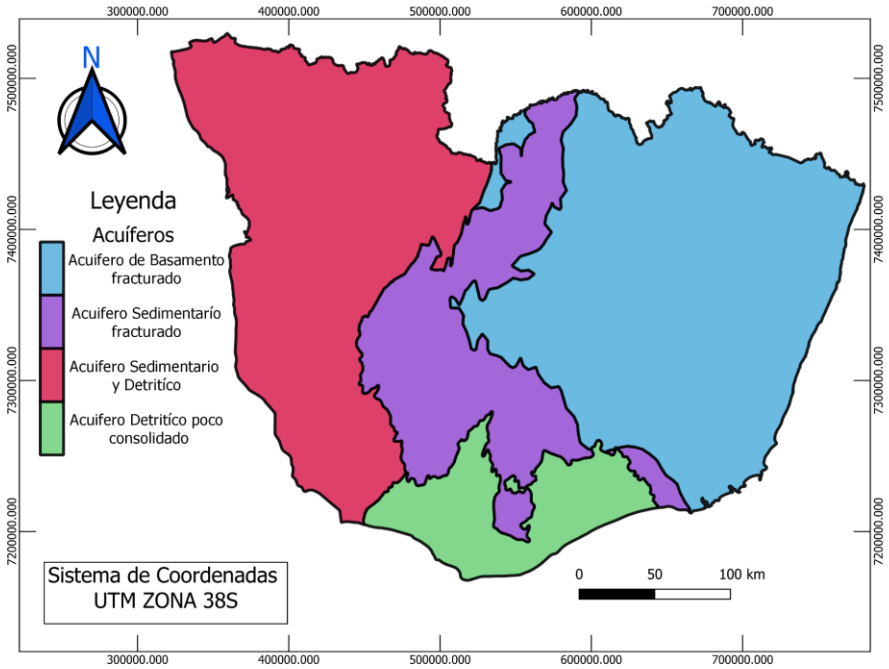

Figura 4: Resumen de las principales tipologías de acuíferos en la zona de estudio (obtenido a través de Al-Zoubaidy et al, 2023).

3.METODOLOGÍA

3.1 Base de datos de pozos

La base de datos utilizada ha sido obtenida a partir del portal de *BushProof* (Al-Zoubaidy M. et al.,2023) e incluye información sobre diversas fuentes, entre ellas, del propio portal de BushProof y también de datos de UNICEF. Esta contiene información de 2449 pozos y sondeos en total, de los cuales se han seleccionado 2314 por estar localizados en la zona de estudio y por incluir información sobre la conductividad eléctrica (CE) de las aguas subterráneas para realizar este trabajo. A su vez, presenta otro tipo de información, como la profundidad del pozo, el nivel freático y si se trata de un punto en funcionamiento, entre otros factores. La conductividad eléctrica presenta un rango de valores muy amplio, siendo el valor más bajo de 7.41 µS/cm y el más alto de 20900 µS/cm. En la Figura 5 se puede observar la distribución espacial de los puntos y como varía la conductividad eléctrica en la zona de estudio. La conductividad eléctrica, que será la variable objetivo a predecir en este trabajo, se define como la capacidad del agua de conducir la corriente eléctrica y esta depende de la cantidad de iones disueltos. Se trata de un parámetro que sirve para cualificar la calidad del agua (Martínez-Santos et al. ,2018).

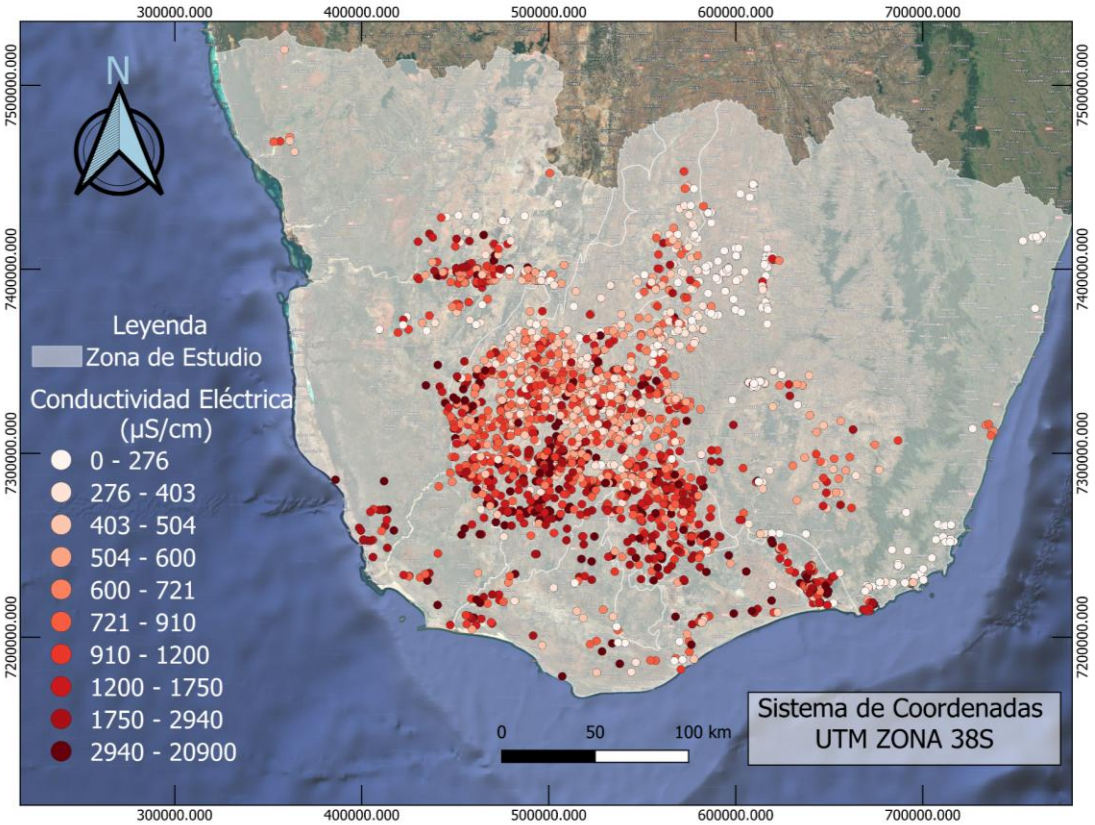

Figura 5: Distribución espacial de los datos de conductividad eléctrica pertenecientes al agua subterránea, medidos en pozos.

3.2 Variables explicativas

La conductividad eléctrica (CE) de las aguas subterráneas depende de diferentes factores, entre los que se encuentran los factores geológicos, climáticos, hidrológicos, topográficos y edafológicos. Por ello, para elaborar las cartografías predictivas de este parámetro, se han recopilado diferentes variables explicativas englobadas en los factores anteriormente mencionados.

Respecto a los factores climáticos se han considerado, la temperatura máxima promedia, la evapotranspiración real, la evapotranspiración media y de Thornthwaite en los periodos anuales húmedos y secos y la precipitación promedia de la zona (Figura6). Estas variables se encuentran altamente relacionadas una con otras y condicionan, en último término, la cantidad de agua que se acaba infiltrando en el acuífero. Dependiendo de la cantidad de agua que se infiltre, la CE puede aumentar o disminuir, ya que, contra menor cantidad de agua haya disponible más concentrada en iones se volverá, aumentando por tanto su CE. Cada uno de estos factores afecta de diferente manera a la conductividad, las temperaturas elevadas provocan un aumento en la evaporación saturando el agua en iones y, por ende, aumentando la CE. Por otro lado, valores más altos de precipitación implican una mayor cantidad de agua, reduciendo así las concentraciones de estos iones y, disminuyendo la CE. Por último, la evapotranspiración se encuentra sujeta a las variables previamente analizadas, en términos generales, se puede asumir que contra más precipitación y temperatura haya, mayor será la tasa de evapotranspiración.

Estas variables fueron obtenidas a través de *Climate Engine*, mediante el Dataset de *Terra Climate* (Abatzoglou et al., 2018; Huntington et al., 2017), el cual resultó ser la mejor opción teniendo en cuenta la resolución espacial de 4km y las variables disponibles.

Para las variables de precipitación, evapotranspiración real media y la temperatura máxima se utilizó el promedio comprendido entre el año 2000 y el 2024, mientras que para la evapotranspiración real y de Thornthwaite se realizaron dos promedios, el primero para un periodo húmedo comprendido entre diciembre y febrero desde el 2000 hasta 2024 y uno seco comprendido entre los meses de julio a septiembre desde el 2000 al 2024. Esta diferenciación de períodos secos y húmedos se pudo realizar gracias a la observación de los datos gráficos anuales de precipitación a través del portal *Terra Climate* (Abatzoglou et al., 2018; Huntington et al., 2017), determinando de este modo que la época más seca o con menos precipitación se corresponde con los meses de julio, agosto y septiembre mientras que la época más húmeda o con más precipitación se corresponde con los meses de diciembre, enero y febrero.

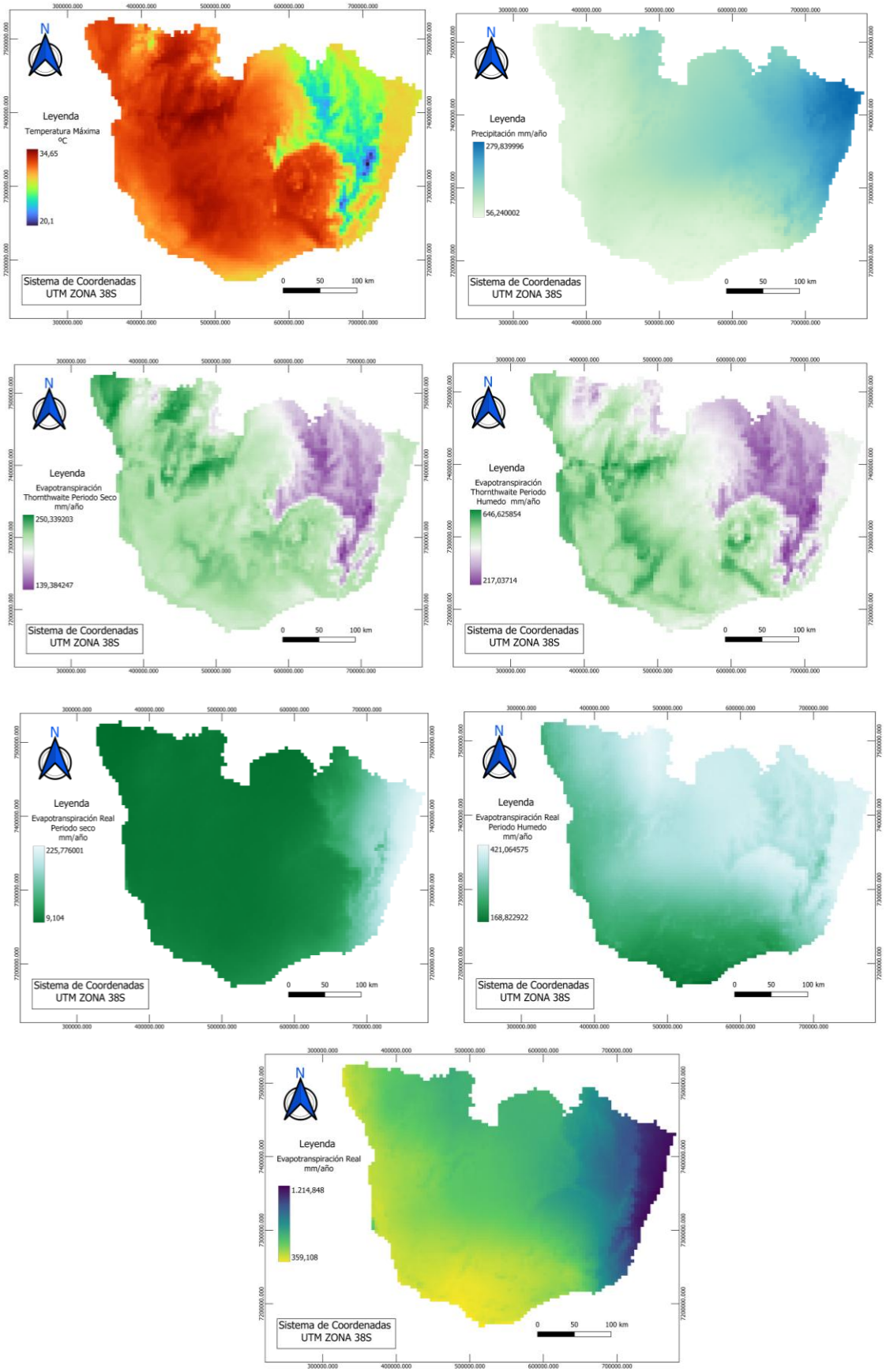

Figura 6:Variables explicativas climáticas para la predicción de CE: Temperatura máxima, Precipitación, evapotranspiración de Thornthwaite para periodos secos y húmedos, Evapotranspiración real y con periodos secos y húmedos.

Los factores hidrológicos suponen una fuente importante de recarga para los acuíferos subterráneos, por lo que condicionan al mismo tiempo la conductividad eléctrica. El agua que se infiltra a través de los ríos, al tratarse de una escorrentía superficial obtenida a través de la precipitación, el agua que se infiltre presentará a priori valores bajos de conductividad eléctrica, mientras que el agua que se pueda infiltrar en la costa a través de la cuña de intrusión marina contendrá una conductividad elevada al presentar grandes concentraciones de sales en disolución, lo que podría tener un alto impacto en los pozos y sondeos más cercanos a la costa. Por ello, se emplearon las siguientes variables: distancia a la costa, distancia a ríos o canales, nivel piezométrico y recarga media anual del acuífero (Figura 7).

Las variables explicativas de distancia a la costa, distancia a ríos y de niveles piezométricos, se descargaron a través del portal de *BushProof* (Al-Zoubaidy M. et al.,2023). Estas capas, en formato *shapefile*, han sido rasterizadas mediante QGIS y, posteriormente, se generaron los mapas de distancias de costa y ríos a partir de la herramienta de QGIS *Proximidad (Distancia raster).* Por otro lado, la recarga media anual del acuífero se obtuvo a través del *British Geological Survey* (MacDonald et al.,2020) en mm/año y se corresponde con los valores estimados para el periodo temporal entre 1970 y 2020.

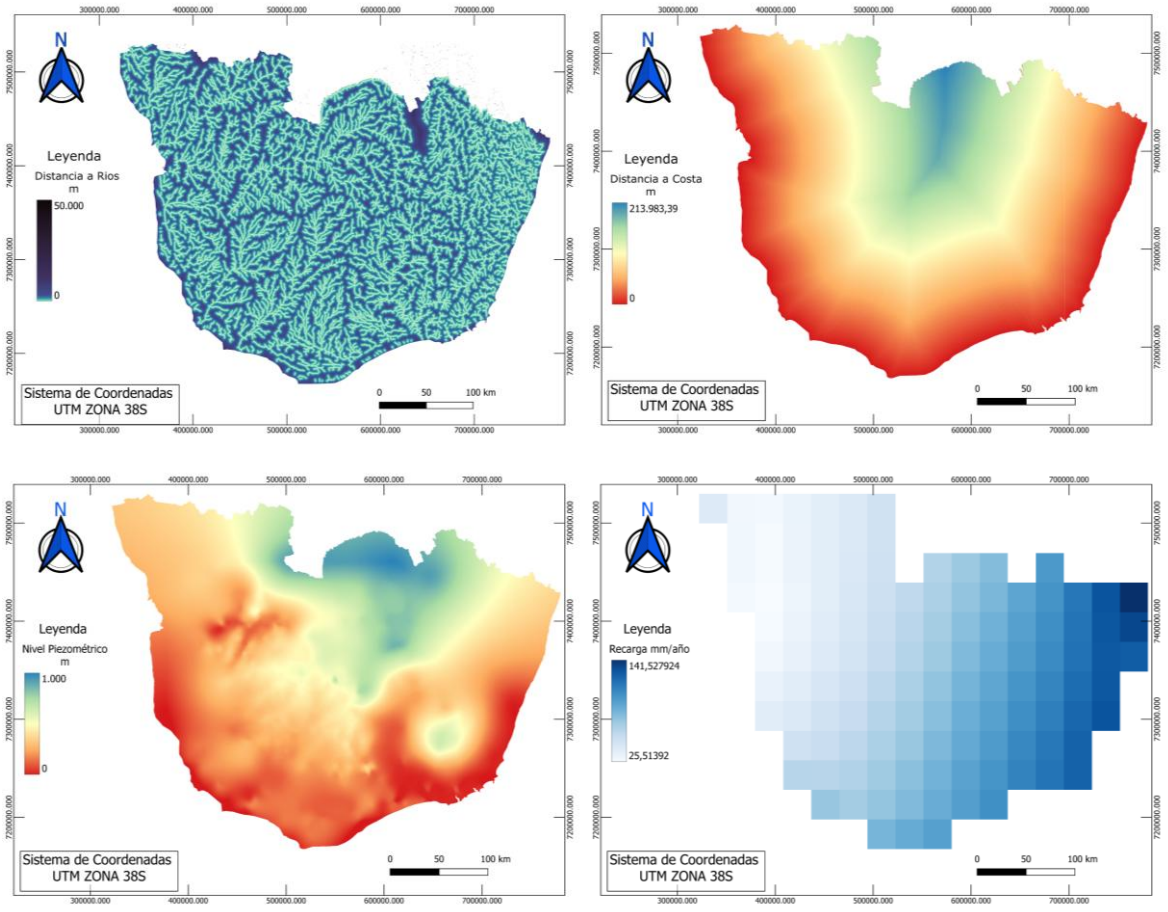

Figura 7: Variables explicativas hidrológicas para la predicción de CE: distancia a los ríos, distancia a la costa, nivel piezométrico y recarga de aguas subterráneas.

Las variables geológicas e hidrogeológicas condicionan de una manera muy importante la conductividad eléctrica de las aguas subterráneas, al conformar el medio por el que circula este recurso. La CE podría aumentar en caso de que el agua atravesará una litología compuesta por rocas evaporíticas, que por lo general son muy susceptibles a la disolución, cargando así de iones y sales el agua. Por otro lado, las características texturales de la roca son también de gran importancia, ya que dependiendo de su porosidad, permeabilidad y fisuración el agua circulará con mayor o menor facilidad, aumentando o reduciendo la velocidad. Si el agua circula más lenta, permanece durante más tiempo en el acuífero, por lo que tiende a aumentar la concentración de iones en disolución aportados por los materiales que conforman el propio acuífero y, por tanto, incrementarla conductividad eléctrica. Otro factor a considerar es la tectónica de la zona, ya que, por ejemplo, en el caso de que haya una falla el agua podría adoptar un flujo preferente por esa fractura, desplazándose más rápido y conectando con diferentes acuíferos, pudiendo provocar una mezcla de aguas. Por ello, se ha tenido en cuenta la distancia a fallas como variable explicativa. El mapa geológico (Figura 8) junto con el mapa de las fallas de la zona se obtuvo en formato *shapefile* a través del portal de *BushProof* (Al-Zoubaidy M. et al.,2023), posteriormente las litologías se separaron en 12 conjuntos, agrupándolas según su edad y tipo de roca. Por su parte, para elaborar la capa de distancia a fallas, se ha empleado un procedimiento similar al empleado previamente con las capas de costa y ríos. En primer lugar, se ha rasterizado la capa y, posteriormente, se ha empleado la herramienta de QGIS *Proximidad (Distancia raster)* para obtenerla distancia a ellas.

También se ha empleado la capa correspondiente al lecho rocoso, obtenida a través del portal *ISRIC World Soil Information* (Hengl, 2015), con una profundidad máxima de 175 cm y una resolución de 250 m (Figura 9).

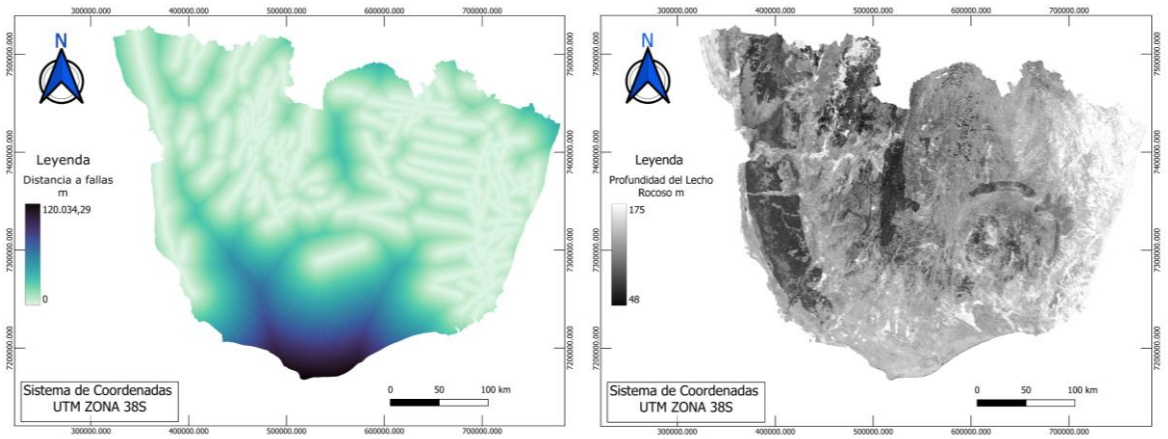

Figura 9:Variables explicativas geológicas para la predicción de CE: Distancia a las fallas y profundidad al lecho rocoso.

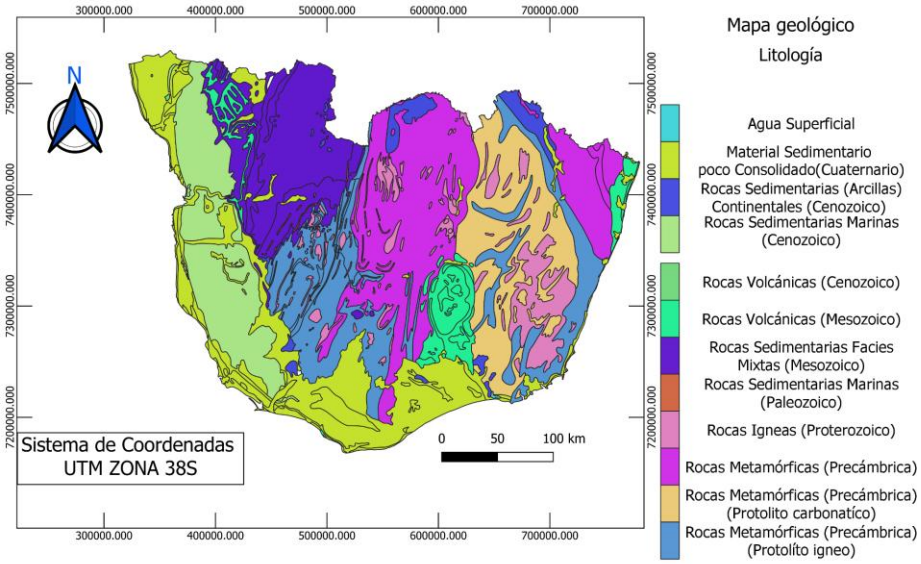

Figura 8:Variables explicativas geológicas para la predicción de CE: Geología

Los factores topográficos condicionan la dirección de flujo del agua, principalmente la componente superficial, provocando que esta se desplace desde un punto de mayor energía a menor energía. Esto puede afectar a la conductividad eléctrica de varias maneras, en zonas de alta pendiente, puede provocar que el agua se desplace más rápido reduciendo así su infiltración hacia el acuífero y manteniéndose en su mayoría como escorrentía superficial. Por el contrario, las pendientes más tendidas tenderán a favorecer la infiltración, aportando una mayor cantidad de agua al acuífero. Por otro lado, la topografía también puede condicionar el espesor saturado del acuífero en ciertas zonas, así como la cercanía del nivel piezométrico a la superficie, en zonas con niveles piezométricos muy someros, donde el agua podría llegar a sufrir evaporación por las altas temperaturas, provocando un aumento de CE en las aguas subterráneas. La topografía se ha obtenido con una resolución de 30 m en formato *raster* (European Space Agency, 2024). Posteriormente, mediante el uso de herramientas de QGIS, se ha obtenido un mapa de pendientes de la zona en grados (Figura 10).

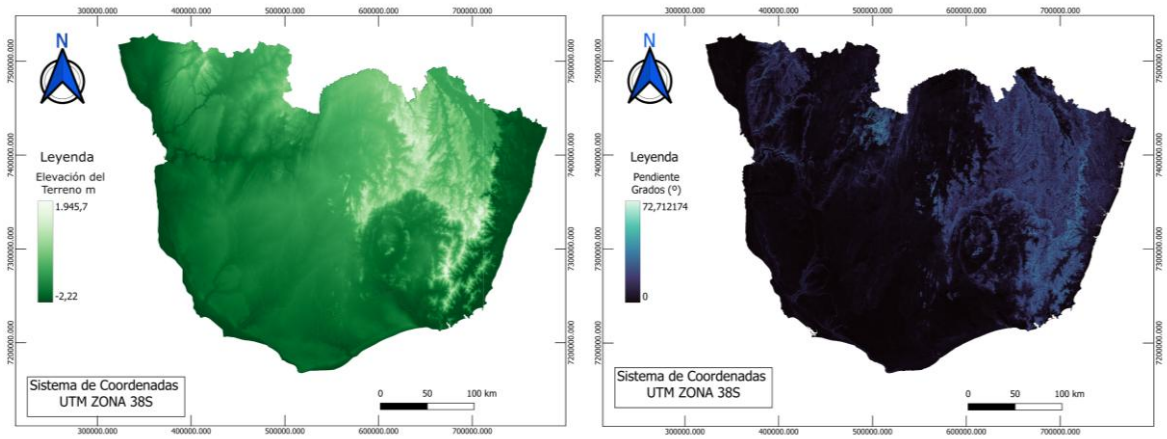

Figura 10: Variables explicativas topográficas para la predicción de CE: elevación del terreno y pendiente.

Por último, se han tenido en cuenta factores más superficiales relacionados con los suelos. Al conformar la primera capa que atraviesa el agua en la infiltración, esta puede modificar de manera importante la CE de las aguas subterráneas. Los suelos con pH más ácido tienden a favorecer la disolución de sales en el agua. A su vez, dependiendo de las características texturales que presente el suelo, el agua podrá infiltrarse más o menos rápido. Por regla general, aquellos suelos que presentan una textura más arenosa permiten una infiltración más rápida, mientras que aquellos que contienen una mayor proporción de arcillas obstaculizan su circulación, aumentando el tiempo de contacto y favoreciendo un posible aumento de CE al adquirir iones pertenecientes a este suelo. Por ello, se han tenido en cuenta cuatro factores principales, el contenido de Arcillas Promedio, el contenido en Arenas Promedio, la Conductividad del suelo Promedio y el pH del suelo Promedio (figura 11). La información necesaria para la elaboración de estas variables fue obtenida a partir del portal *ISRIC World Soil Information*. Las capas, en formato *raster*, presentan una resolución de 250 m y los porcentajes y datos se encuentran disponibles a diferentes profundidades (0 a 5cm, de 5 a 15cm, de 15 a 30cm, de 30 a 60cm, de 60 a 100cm y de 100 a 200 cm). Posteriormente, empleando la *calculadora raster* perteneciente a QGIS se ha calculado el contenido promedio con todas las profundidades para cada una de las variables.

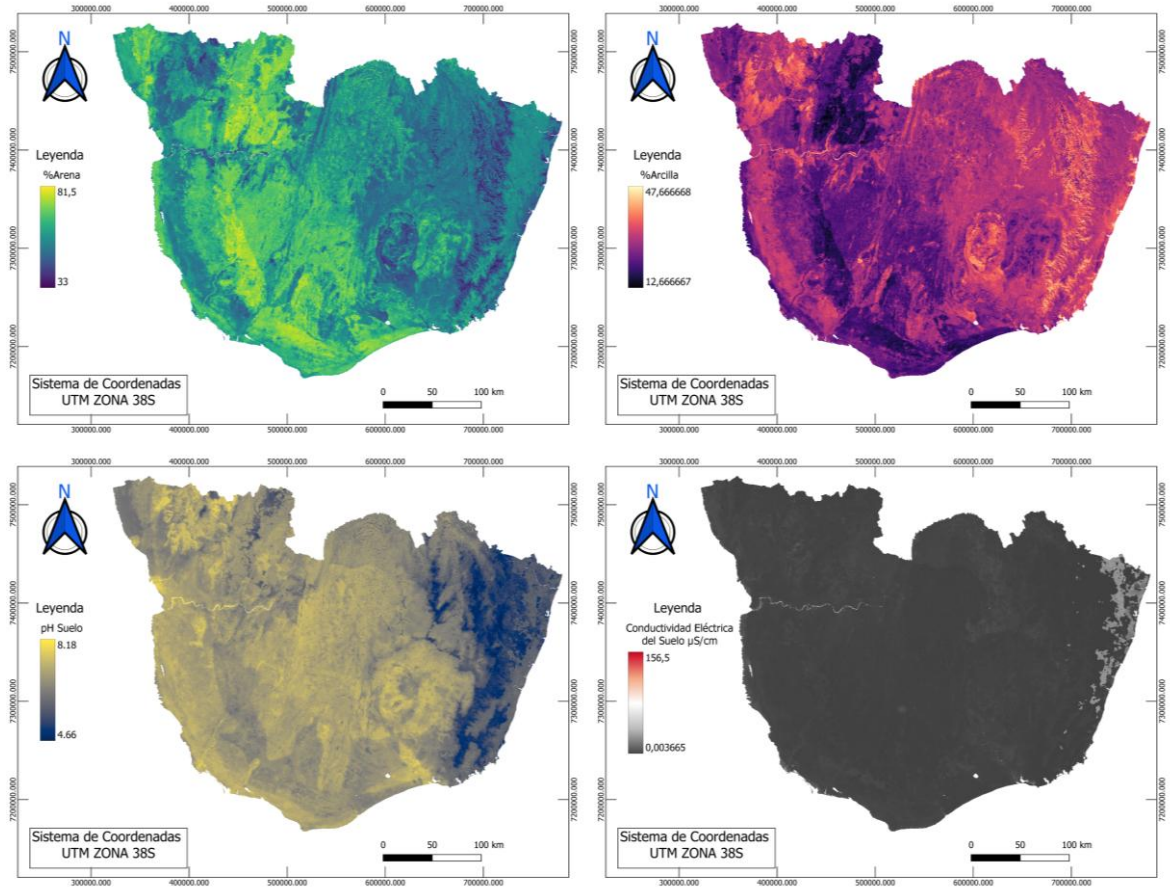

Figura 11: Variables explicativas superficiales para la predicción de CE: % de Arenas, % de Arcillas, pH del suelo y conductividad eléctrica del suelo.

A continuación, se muestra un resumen de todas las variables explicativas utilizadas, indicando sus unidades, resolución, periodo temporal y la base de datos de origen (Tabla 1).

Tabla 1: Total de variables y sus correspondientes unidades, abreviaturas, resolución, periodo temporal y la fuente de donde se ha obtenido cada una.

	Variable	Abreviaturas	Unidad	Resolución	Periodo Temporal	Fuente
Factores Climáticos	Evapotranspiración real media	Evap.R.	mm/año	4km	01/01/2000-31/12/2024	Terra Climate - Climate Engine
	Evapotranspiración real estación seca	M.Evap.R.S.	mm/año	4km	Julio a Septiembre 2000-2024	Terra Climate - Climate Engine
	Evapotranspiración real estación humeda	M.Evap.R.H.	mm/año	4km	Diciembre a Febrero 2000-2024	Terra Climate - Climate Engine
	Evapotranspiración Thornthwaite estación seca	M.Thorn.S.	mm/año	4km	Julio a Septiembre 2000-2024	Terra Climate - Climate Engine
	Evapotranspiración Thornthwaite estación humeda	M.Thorn.H.	mm/año	4km	Diciembre a Febrero 2000-2024	Terra Climate - Climate Engine
	Precitación promedia	PPT	mm/año	4km	01/01/2000-31/12/2024	Terra Climate - Climate Engine
	Temperatura máxima promedia	Temperatura Max	ºC	4km	01/01/2000-31/12/2024	Terra Climate - Climate Engine
Factores Superficiales	Contenido en Arcillas Promedio	% Arcillas	g/kg	250m	-	ISRIC- World Soil Information
	Contenido en Arenas Promedio	% Arenas	g/kg	250m	-	ISRIC- World Soil Information
	Conductividad del suelo	Conductividad S.	µS/cm	250m	-	ISRIC- World Soil Information
	pH del suelo	pH S.	-	250m	-	ISRIC- World Soil Information
Factores Topográficos	Modelo Digital de Elevación	Topo	m	30 m	-	European Space Agency
	Pendientes	Pendiente	º	31 m	-	Elaborado a partir del DEM
Factores Hidrogeológicos	Distancia a la Costa	Dis.Costa	m	30m	-	Elaborado a partir del DEM
	Distancia de Los Rios o Canales	Dis.Rios	m	30m	-	Elaborado a partir del DEM
	Recarga del Acuifero	Recarga	mm/año	32km	01/01/1970-01/01/2020	BGS (Groundwater recharge in Africa from ground based measurements)
	Niveles Piezométricos	Isolinias	m	30m	-	BushProof y Elavorado a partir de DEM
	Altura del Lecho Rocoso	Lecho	m	250m	01/01/2011-10/01/2015	ISRIC- World Soil Information
Factores Geológicos	Litología	Geo	-	-	-	BushProof
	Distancia a Fallas	Dis.Fallas	-	30m	-	BushProof y Elaborado a partir del DEM

3.3 Métodos de aprendizaje automático (machine learning)

El aprendizaje automático (*machine learning en inglés),* forma parte de un sector dentro de la inteligencia artificial, el cual plantea entrenar a los ordenadores mediante algoritmos matemáticos para que sean capaces de relacionar y aprender a partir de datos, es decir, que sean capaces de aprender de forma automática (Géron, 2019; Gómez-Escalonilla, 2024). A la hora de trabajar con esta serie de algoritmos cabe diferenciar dos tipos de variables que se pueden usar en ellos. La primera sería la variable objetivo, o también denominada variable dependiente. Esta variable será aquella que se quiera predecir mediante los algoritmos de aprendizaje automático, en este caso se trata de la conductividad eléctrica. Por otro lado, se emplean las variables explicativas, también denominadas variables independientes, estas

variables serán las que tratarán de explicar o describir la variable objetivo, haciendo referencia a los factores vistos anteriormente.

Existen dos enfoques *machine learning* dependiendo del objetivo y de la base de datos, aprendizaje no supervisado y aprendizaje supervisado (Gómez-Escalonilla, 2024). Los algoritmos de aprendizaje no supervisado son aquellos que no tienen una variable objetivo, es decir, no tienen como finalidad obtener una predicción de un factor en concreto, sino que tratan de extraer información de los diversos conjuntos de datos o agruparlos por similitud según las métricas obtenidas (Gómez-Escalonilla, 2024). Por otro lado, los algoritmos de aprendizaje supervisado son aquellos que necesitan una base de datos etiquetada, en otras palabras, necesita conocer tanto los valores de las variables explicativas como de la variable objetivo, de este modo se podrán detectar patrones entre las variables explicativas que nos permitan predecir la variable objetivo (Géron, 2019; Suthaharan, 2016).

Dentro de los algoritmos supervisados se pueden diferenciar dos tipos, siendo estos de regresión o de clasificación. Los algoritmos de regresión se caracterizan por aprender y predecir una variable objetivo de tipo continuo, como, por ejemplo, los valores de evolución de un caudal de pozo de extracción. Por su parte, los algoritmos de clasificación son entrenados y se usan para predecir una variable objetivo categórica o discreta, lo que significa que van a predecir entre diferentes clases, pudiendo ser binario, prediciendo solo dos clases (p.ej pozo negativo o positivo) o multiclase, en la cual se tendrán en cuenta una clasificación con un mayor número de categorías (Gómez-Escalonilla, 2024).

A la hora de realizar este trabajo, se han tenido en cuenta como datos de entrada tanto variables explicativas como la variable objetivo, por ello el enfoque empleado será el de un algoritmo supervisado. Además, la variable objetivo-relacionada con la conductividad eléctrica, ha sido clasificada con un procedimiento de carácter binario, ordenando los valores dependiendo de si superan o no un umbral establecido de conductividad eléctrica, siendo el valor 0 para aquellos que lo superen y el valor 1 para aquellos que se encuentren por debajo de ese límite, siendo por tanto aptos siguiendo ese umbral.

En un primer momento, se valoraron cuatro umbrales diferentes de conductividad: 500 µS/cm, 1000 µS/cm, 2000 µS/cm y 3000 µS/cm. Sin embargo, el valor de 3000 µS/cm fue descartado por varios factores. En primer lugar, los puntos de agua que superan este umbral suponen menos del 10% del total, lo que podría afectar negativamente a la hora de ejecutar los algoritmos de *machine learning*. Por otro lado, el límite de 3000 µS/cm superaba el rango de salinidad máxima establecida para el consumo humano, estando el límite en 2700 µS/cm, por lo que no podría servir como un límite máximo de CE (WHO, 2021). Por ello, en este trabajo se emplearon solo los rangos de 500 µS/cm, 1000 µS/cm y 2000 µS/cm, siendo los de 500

µS/cm y 1000 µS/cm los valores estándar para el agua de buena calidad y sabor aceptable y el de 2000 µS/cm como agua de calidad estándar (WHO, 2021) (Tabla 2)

Tabla 2: Cantidad de puntos aptos y no aptos para cada uno de los umbrales de CE tomados en consideración.

Valores	500 µS/cm		1000 µS/cm		2000 µS/cm		3000 µS/cm	
	Aptos (1)	No Aptos (0)	Aptos (1)	No Aptos (0)	Aptos (1)	No Aptos (0)	Aptos (1)	No Aptos (0)
Puntos	772	1677	1589	860	2063	386	2225	224
%	31.52%	68.48%	64.88%	35.12%	84.24%	15.76%	90.85%	9.15%

3.4 Software empleado

El software empleado para la elaboración de las cartografías predictivas MLMapper V2.0. Un código programado en Python, utilizado como complemento para la herramienta de QGIS (Gómez Escalonilla et al., 2022). Esta herramienta implementa diferentes procedimientos estándar para el tratamiento de los datos, incluyendo el análisis de multicolinealidad, técnicas de escalado de las variables explicativas y la optimización de los hiperparámetros mediante búsqueda aleatoria (Gómez-Escalonilla, 2024).

MLMapper incluye 15 algoritmos de clasificación supervisada diferentes. entre ellos se encuentran modelos de tipo *ensemble*, basados en árboles, como el *Gradient Boosting Classifier* (GBC), el *Random Forest Classifier* (RFC) y *Extra Trees Classifier* (ETC), así como otros basados en redes neuronales (MLP), en regresión logística (LRP), en clasificador k-vecinos (KNN) y en clasificador de vectores de soporte (SVC), entre otros (Gómez-Escalonilla et al., 2022).

Al trabajar con este software lo primero que se debe realizar es una base de datos en formato CSV que incluya la información de los pozos de agua, estando esta información colocada por columnas con un orden específico, siendo la primera columna la variable objetivo ya clasificada mediante los umbrales previamente descritos, con valores de 0 o 1, posteriormente las coordenadas de los pozos (X e Y) y, a continuación, los valores de las 20 variables explicativas que se van a utilizar. Por otro lado, es necesario generar otro archivo para llevar a cabo la cartografía predictiva. Se trata de una malla regular de puntos regulares distanciados entre sí por 1000 metros. Este archivo debe contener tanto las coordenadas de los puntos como los valores de las variables explicativas, ordenándose de manera idéntica al archivo previamente descrito, ya que, en caso de que no se encuentren clasificados con el mismo orden se producirá un error y no se podrá ejecutar el programa. Para asignar los valores necesarios de las variables explicativas a cada uno de los puntos se ha empleado la herramienta *Point Sampling* en QGIS. Esta se encarga de combinar la información de cada pozo con los valores respectivos a las capas de las variables explicativas en esa precisa localización.

Una vez obtenido el CSV (con los datos separados por comas), este deberá pasar por un preprocesamiento previo a la ejecución de MLMapper. Este procedimiento comprime los valores de las variables explicativas facilitando las tareas de los algoritmos, ya que en muchas ocasiones algunos de los rangos numéricos presentan valores mucho más elevados que otras variables, lo que puede condicionar a los modelos programa erróneamente a pensar que las variables con valores más altos son más importantes o tienen más peso que las demás, dejando de lado a otras que podrían llegar a ser igual o más relevantes. Para realizar este preprocesamiento, se ha usado el escalador MaxAbs (*Maximal absolute scaler*), el cual se encarga de normalizar todos los datos en un rango de [-1,1], dividiendo los valores de cada variable por el valor más alto de cada una (Gómez-Escalonilla, 2024; Pedregosa et al., 2011; Zheng & Casari, 2018).

Los algoritmos pueden atribuir un peso extra a aquellas variables explicativas que presenten una alta correlación, lo que puede generar problemas de multicolinealidad. Para analizarlo, se ha utilizado el coeficiente de correlación de Pearson, que mide el grado de covariación entre las variables clasificándolas entre los valores de 1 y -1. Los valores positivos indican una correlación directa y los negativos inversa, contra más cercanos sean los valores al 1 y -1 más correlación habrá entre el par de variables analizadas, mientras que contra más se acerquen al 0 menor será está correlación. Por ello, las variables que presenten una alta correlación, al resultar redundantes, pueden llegar a generar una problemática en la fase de aprendizaje. Sin embargo, siempre hay que tener en cuenta la importancia de las variables antes de eliminarlas, ya que en ocasiones pueden aportar información indispensable para la elaboración de la cartografía predictiva.

Otro aspecto a destacar en los enfoques de clasificación supervisada es que, en ocasiones, los algoritmos tienden a infravalorar aquella información o datos que se encuentran en una minoría, es decir, se fijan en la clase mayoritaria. En este trabajo bajo ninguno de los umbrales establecidos para la CE, se llega a obtener un equilibrio exacto entre puntos aptos (positivos) y no aptos (negativos), provocando que el extremo con mayor número de puntos predomine en la fase de entrenamiento. Por ello, con el objetivo de lograr un equilibrio entre ambas clases en cada uno de los escenarios, se ha optado por aplicar técnicas de sobremuestreo, en concreto la técnica SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002). Esta técnica crea nuevos puntos sintéticos de la clase minoritaria de forma aleatoria utilizando como referencia el espacio multidimensional (con tantas dimensiones como variables explicativas se utilicen= de las muestras. Para ello, genera estas muestras sintéticas sobre las líneas que unirían a los diferentes vecinos de la clase minoritaria y adquiriendo un valor equivalente intermedio para todas las variables explicativas. Es importante destacar, que

este procedimiento se ha empleado únicamente en el conjunto de datos de entrenamiento para evitar problemas de “fuga de datos” o *data leakage* (Kaufman et al., 2012).

A continuación, se procederá a realizar una división aleatoria de puntos o pozos dentro del programa de MLMapper. Un porcentaje de ellos servirá como entrenamiento de los algoritmos y el porcentaje restante, se empleará como validación de los modelos. El conjunto dedicado al entrenamiento servirá para que los algoritmos traten de aprender los patrones entre las variables explicativas y la variable objeto, es decir la CE (Gómez-Escalonilla, 2024). Para este trabajo se han utilizado cuatro porcentajes diferentes de entrenamiento para cada umbral de conductividad eléctrica, siendo estos el 60%, 70%, 80% y 90%, dando lugar a un total de 12 resultados o escenarios diferentes. Esto se ha realizado para profundizar en el análisis y observar qué parámetro es el más adecuado u obtiene mejores resultados a la hora de predecir la CE. Cabe destacar, que, durante la etapa de entrenamiento, los algoritmos han sido optimizados mediante un procedimiento de Búsqueda Aleatoria (*Random Search procedure*). Este consiste en ajustar los hiperparámetros de los algoritmos en búsqueda de la mejor arquitectura teniendo en cuenta la base de datos empleada. Se han aplicado 50 iteraciones para esta optimización, es decir, se han evaluado 50 combinaciones de hiperparámetros diferentes en búsqueda de la más óptima.

El conjunto de validación se corresponde con el porcentaje inverso al de entrenamiento, es decir, si se utiliza un 60% de los pozos como porcentaje de entrenamiento, el 40% restante será empleado para la validación. Esta etapa permitirá evaluar la eficacia de predicción del algoritmo, ya que, el algoritmo tendrá que pronosticar sus valores mediante los patrones entre la variable objetivo y las diversas variables explicativas aprendidas en la fase de entrenamiento.
Una vez ejecutado el programa se obtendrán una serie de métricas, las cuales indicarán el porcentaje de aciertos obtenidos por cada método de análisis a la hora de predecir el valor de la conductividad en los pozos utilizados para la validación, esto permite comprobar la fiabilidad de las predicciones, así como su posible error, tanto para aquellos pozos que se encuentren por encima del umbral de la variable objetivo como para los que se encuentren por debajo. De este modo también se podrá estipular cuál de los métodos de análisis es el más adecuado, ya que en un principio no se podía determinar cuál de ellos tendría mayor compatibilidad con la base de datos aportada.

En este trabajo se ha empleado un enfoque de clasificación binaria. Teniendo esto en cuenta, para determinar el rendimiento de los algoritmos se ha hecho uso de diferentes métricas estándar empleadas en este tipo de trabajos. La primera consiste en una matriz de confusión, que resume la cantidad de aciertos y errores del modelo, clasificándolos en 4 categorías diferentes: verdaderos negativos (TN), verdaderos positivos (TP), falsos negativos (FN) y falsos positivos (FP). Los verdaderos (ya sean positivos o negativos) serán puntos en los que los algoritmos han encontrado una serie de patrones de la variable explicativa que los ha llevado a predecir de forma correcta la variable objetivo. Por el contrario, los falsos positivos o negativos son puntos que los algoritmos han predicho de forma incorrecta.
A partir de estos datos se han obtenido una serie de métricas que se explicarán a continuación: *Test score*, precisión, *Recall, F1-score* y *AUC ROC.*
La primera es el *Test Score*, la cual equivale a la proporción total de aciertos en la predicción, siendo 1 el máximo de aciertos posible y 0 ninguno, indicando una mayor capacidad predictiva contra más se acerque el número a 1 (Gómez-Escalonilla, 2024).

Otro factor a tener en cuenta es la precisión, la cual indica con exactitud el porcentaje de predicciones verdaderamente positivas y negativas.
La métrica *Recall* (sensibilidad) tiene la capacidad de detectar correctamente tanto los verdaderos positivos (TP) de los positivos reales (TP+FN), como de los verdaderos negativos (TN) de los negativos reales (TN+FP), obteniendo la tasa de verdaderos positivos y negativos.
El *F1-score* (puntuación F1) representa la media armónica entre la precisión y *recall*, indicando la cantidad de puntos aptos o no aptos se han conseguido predecir de cada uno.
Para finalizar se tendrá en cuenta la métrica de *AUC ROC* (*Area Under Receiver Operating Characteristic Curve*), la cual mide la capacidad del modelo para discriminar entre clases. en esta métrica el valor máximo es 1 y el mínimo 0, un valor de 0.5 de sugiere que no hay discriminación entre clases. Según la literatura, valores entre 0.7 a 0.8 se consideran aceptables y por encima de 0.8 son equivalente a una excelente capacidad predictiva (Hosmer and Lemeshow, 2000).

Si los modelos obtienen unos buenos resultados en cuanto a métricas que evalúan su capacidad predictiva, se asume que pueden predecir la variable objetivo CE. Por lo tanto, se procederá a la obtención de las predicciones para cada punto de la malla anteriormente elaborada mediante dos tipos diferentes de pronóstico, uno binario, el cual indica si el pozo supera o no el umbral establecido (1 o 0) y otro mediante probabilidades, estimando la probabilidad que tiene ese punto de estar por encima o por debajo del umbral de la variable objetivo.

A continuación, se muestra un resumen del proceso metodológico (Figura 12).

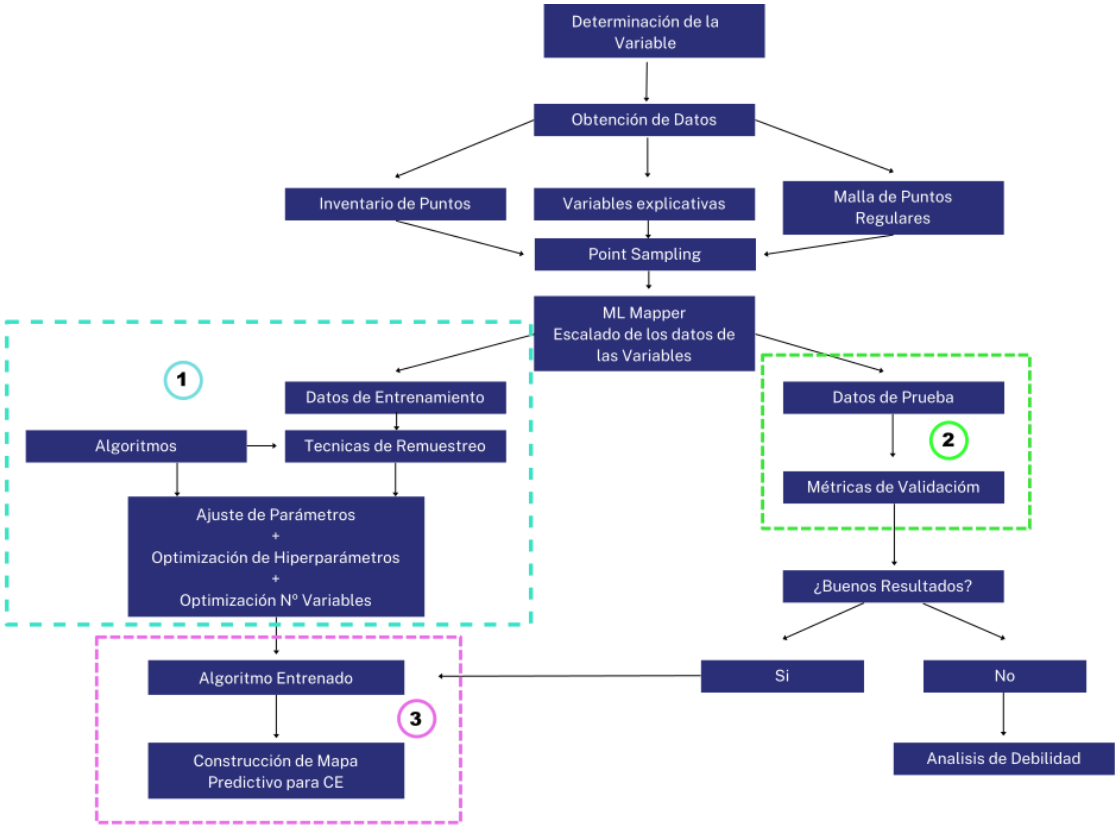

Figura 12: Esquema de funcionamiento de MLMapper V2.0 aplicado en la predicción espacial de la conductividad eléctrica de las aguas subterráneas en donde: 1 es la fase de entrenamiento, 2 es la fase de validación y 3 es la fase de predicción (Modificado de Gómez-Escalonilla, 2024).

3.5 Estimación de población en riesgo

Otro de los objetivos del trabajo consiste en estimar la población en riesgo, por estar localizada en zonas donde las predicciones arrojan una alta probabilidad, de superar el umbral de conductividad eléctrica establecido. Tras obtener las métricas de los diferentes escenarios, se ha seleccionado aquel con mejores resultados junto con los algoritmos de mejor rendimiento, realizando con ellos una cartografía predictiva en formato raster de la probabilidad que habría de superar el umbral determinado de CE. Posteriormente, mediante la herramienta QGIS se realizó un análisis que permitió estimar la población en posible riesgo de consumo de aguas con excesiva salinidad. Para ello, se descargaron a través del portal *The Humanitarian data exchange*, tanto los límites poblacionales, como los datos de población de las distintas unidades administrativas de la zona de estudio (OCHA, 2018).

4.RESULTADOS Y DISCUSIÓN

4.1. Matriz de correlación

El análisis de multicolinealidad de las 20 variables se representa a partir de la matriz de correlación (Figura 13). Las correlaciones directas están representadas en color azul, y las correlaciones inversas se muestran en color rojo. Colores más intensos, rojo o azul, indican una mayor correlación (cercanas a +1 o -1) entre el par de variables.

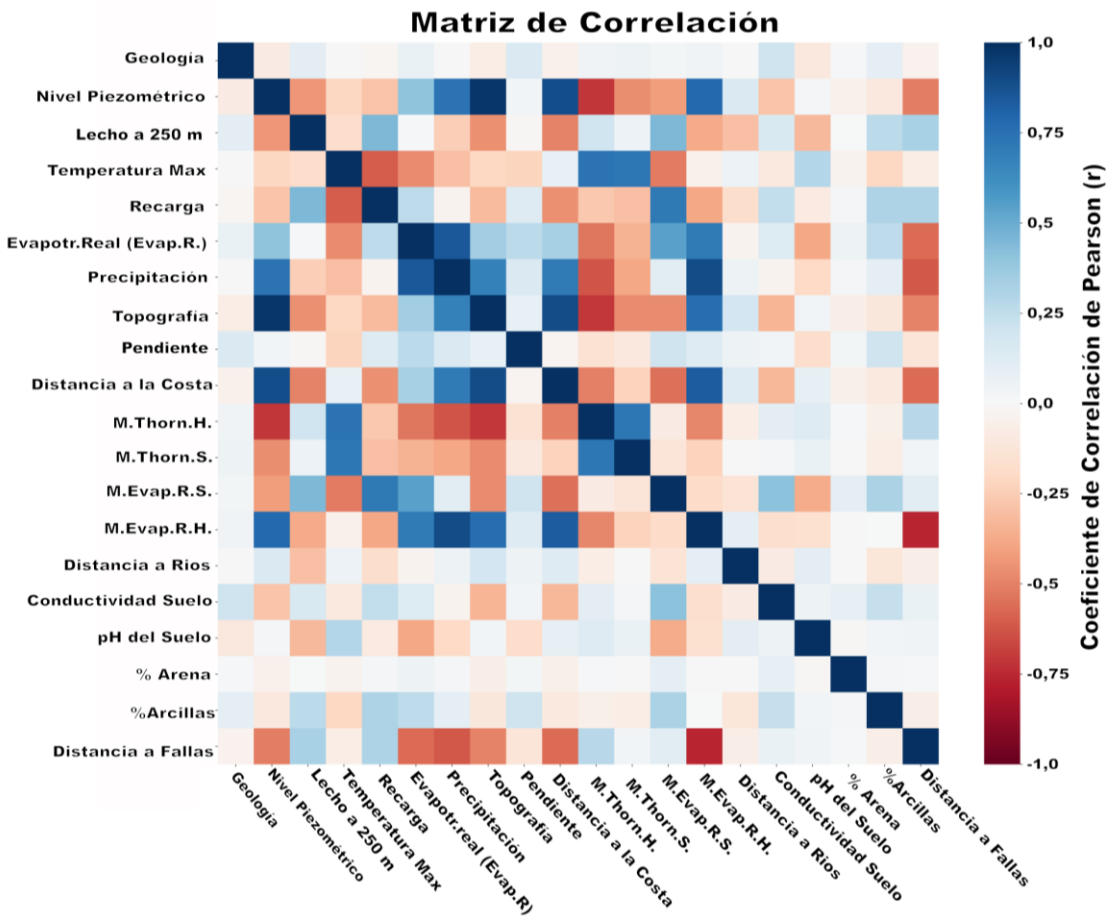

Figura 13: Matriz de correlación entre las diferentes variables explicativas empleadas

Dentro de la matriz se pueden observar fuertes correlaciones entre varias variables explicativas, tanto directas (valores de Pearson > 0.75) como inversas (valores de Pearson < -0.75), tal y como se puede observar en la Tabla 3.

Tabla 3: Correlaciones entre variables que presentan un alto nivel de multicolinealidad

Correlación		
Directas		
Nivel Piezométrico	-	Precipitación
Nivel Piezométrico	-	Topografía
Nivel Piezométrico	-	Distancia a la Costa
Evapotranspiración	-	Precipitación
Distancia a la costa	-	Topografía
M.Evapo.R.H.	-	Precipitación
M.Evapo.R.H.	-	Distancia a la Costa
Inversas		
Nivel Piezométrico	-	M.Thorn.H.
Topografía	-	M.Thorn.H.
Distancia Fallas	-	M.Evapo.R.H.

Algunas de estas correlaciones eran previsibles desde el inicio. En los primeros casos, el nivel piezométrico está altamente condicionado por la precipitación, ya que es la que aporta el agua de recarga, por lo que a priori una mayor precipitación estaría correlacionada con una mayor cota del nivel piezométrico.

Por otro lado, también es esperable que una mayor altitud topográfica esté relacionada con valores más altos de cota del nivel piezométrico, ya que, como se ha podido observar hay una mayor precipitación en las zonas altas, aportando agua al acuífero y elevando su nivel piezométrico y, además, en determinados contextos geológicos, el nivel piezométrico replica en gran medida la topografía de superficie.

Respecto a la correlación directa entre el nivel piezométrico y la distancia a la costa esta se justifica porque el nivel piezométrico desciende junto a la topografía hasta el punto de intersección con el nivel del mar.

En el caso de la evapotranspiración real, tanto en los meses húmedos como anualmente, la precipitación estará directamente correlacionada con estas variables, puesto que contra más precipitación haya, más cantidad de agua hay disponible para su evapotranspiración, en otras palabras, si no hay precipitación no hay evapotranspiración.

Respecto a la correlación de la evapotranspiración real en los meses húmedos con la distancia de la costa, se debe primordialmente a la temperatura. Se pudo observar que las temperaturas son más elevadas en la costa, sobre todo la costa oeste de la zona. Una elevada temperatura provoca una mayor tasa de evapotranspiración y, por ello, contra más cerca esté de la costa mayor evapotranspiración habrá.

En el caso de las correlaciones inversas, es lógico que mayores tasas de evapotranspiración, que limitan la infiltración del agua y la recarga del acuífero, estén correlacionadas espacialmente con cotas más deprimidas del nivel piezométrico.

Respecto a la correlación inversa entre la evapotranspiración de Thornthwaite en el periodo húmedo y la topografía está afectada principalmente por la temperatura. Al ser esta última

variable más alta en zonas deprimidas topográficamente, va a provocar un aumento de las tasas de evapotranspiración.
Por último, en cuanto a la distancias a las fallas, no se ha podido estipular una causa clara entre la relación inversa con la evapotranspiración real en periodos húmedos.
Aunque las correlaciones entre las variables explicativas indicadas anteriormente presentan valores elevados, estas contienen información de interés para obtener la cartografía predictiva y algunas de ellas, provienen de fuentes muy diferentes, por ello, no se las ha considerado redundantes y se han mantenido en las fases posteriores de la investigación.

4.2. Evaluación de los algoritmos

En la etapa inicial, se utilizó la herramienta MLMapper sin incluir la aplicación de la técnica de sobremuestreo SMOTE. Sin embargo, los resultados obtenidos mostraron una tendencia a la sobrepredicción de la clase mayoritaria frente a la minoritaria, puesto que, a pesar de mostrar métricas elevadas en términos generales (*Test Score*), los algoritmos mostraban valores *F1-score* muy diferentes para ambas clases. En este caso, el porcentaje de aciertos en los puntos positivos o aptos era mucho más alto que en los puntos no aptos o negativos y viceversa, dependiendo del umbral. Esta problemática tiende a aparecer cuando las bases de datos presentan un desequilibrio entre clases, como es el caso de la empleada en este trabajo. Por ello, los resultados que se muestran a continuación hacen referencia a los obtenidos únicamente tras la aplicación de la técnica SMOTE para garantizar el equilibrio entre clases.
Por ello, los resultados de este trabajo se han centrado en profundizar en el análisis del impacto que tienen tanto los umbrales seleccionados para discriminar entre apto y no apto, como las divisiones *train/test* empleados por los algoritmos.

A continuación, se muestran las métricas de evaluación obtenidas por los algoritmos con mejor rendimiento para cada uno de los 3 umbrales de CE evaluados, es decir, 500 µS/cm, 1000 µS/cm y 2000 µS/cm (Tablas 4, 5, 6) Para cada uno de estos umbrales se han seleccionado los 3 algoritmos que presentaban una mayor capacidad predictiva para cada porcentaje de *train/test*, de este modo se puede observar en su conjunto, cuáles son los algoritmos que mejor funcionan para la base de datos empleada en este trabajo.
Para el umbral de 500 µS/cm (Tabla 4), los algoritmos con mejores resultados son *Gradient Boosting Classifier* (GBC), *Support Vector Machines* (SVC), *Logistic Regression (LRG), Random Forest Classifier (RFC), K-Neighbors (KNN), Extra Trees Classifier* (ETC) y *Quadratic Discriminant Analysis* (QDA), destacando el algoritmo *Gradient Boosting Classifier* (GBC) por obtener buenos resultados en todos las divisiones *train/test* y los algoritmos de SVC y LRG que presentan también buenos resultados, con métricas óptimas en la mitad de

los escenarios de *train/test.* En todos los casos, el *test score* promedio supera el valor de 0.73 y el AUC el 0.75, obteniendo hasta un 0.8. Las divisiones t*rain/test* con mejores *test score* promedio son las de 0.6 y 0.9, con valores de *test score* promedio de 0.74 y 0.76 respectivamente, indicando que los valores extremos del *train/test* serían los mejores al predecir los puntos negativos y positivos. Respecto a la métrica *F1-score*, se observan valores más altos en las clases negativas (0 o no aptos) en comparación con las positivas (1 o aptos). Los *F1-score* (0) presentan valores superiores a 0.79 mientras que los valores *F1-score* (1) no superan el 0.63 siendo su valor más bajo de 0.58. Esto indica que los algoritmos son capaces de distinguir mejor los puntos no aptos o negativos que los positivos, esto seguramente se deba a que los algoritmos disponen de un un mayor porcentaje de puntos de esa clase durante la fase de entrenamiento. Las divisiones t*rain/test* con mejores métricas generales de *F1- score* serían los que emplean un 80% y 90% de los datos para entrenamiento, indicando que contra más puntos se usen en la fase de entrenamiento mejor consigue diferenciar y predecir el programa los aptos de no aptos de las muestras utilizadas para el test.

Tabla 4: Resultados obtenidos de las métricas de evaluación para el umbral de 500 µS/cm para los diferentes test/score.

Umbral 500 µS/cm					
Train/Test	**Mejores modelos**	**Test score promedio**	**F1 score 0 promedio**	**F1 score 1 promedio**	**AUC promedio**
0.6/0.4	Support Vector Machines Random Forest Classifier Gradient Boosting Classifier	0.74	0.81	0.58	0.78
0.7/0.3	Gradient Boosting Classifier Support Vector Machines K-Neightbors	0.74	0.81	0.57	0.75
0.8/0.2	Gradient Boosting Classifier Extra Trees Classifier Logistic Regress	0.73	0.79	0.62	0.79
0.9/0.1	Gradient Boosting Classifier Logistic Regress Quadratic Discriminant Analysis	0.76	0.82	0.63	0.80

En el escenario que emplea el umbral de 1000 µS/cm para discriminar entre puntos positivos (aptos) y negativos (no aptos) (Tabla 5), los algoritmos con mejores resultados son *Gradient Boosting Classifier* (GBC), *Random Forest Classifier (RFC)*, *Extra Trees Classifier* (ETC), *Support Vector Machines* (SVC), *Logistic Regresión* (LRG) y *K-neighbors* (KNN). Destacan

los algoritmos *ensemble* basados en árboles, que incluyen los modelos *Gradient Boosting Classifier* (RBC), *Random Forest Classifier (RFC)* y *Extra Trees Classifier* (ETC), que presentan métricas elevadas en más de la mitad de los escenarios de divisiones de *train/test.* El *test score* promedio supera los 0.71 y el AUC el 0.75, obteniendo hasta un 0.78. Las divisiones de *train/test* con mejores resultados para la métrica *test score* son las que emplean un 60% y 70% de los datos para entrenamiento, con valores de 0.72 para ambos casos. Respecto a la métrica *F1-score*, se observan valores más altos en las clases positivas (1 o aptos) en comparación con las negativas (0 o no aptos). Los *F1-score* (1) presentan valores superiores a 0.76 mientras que los valores *F1-score* (0) no superan el 0,66 siendo su valor más bajo de 0.59, indicando que los algoritmos distinguen de manera óptima entre las clases negativas y positivas, con pequeñas diferencias de rendimiento a la hora de predecir ambas clases. Las divisiones de *train/test* con mejores métricas en el conjunto global son 0.6 y 0.7. Atendiendo a los resultados obtenidos en este umbral de CE, un menor porcentaje de puntos empleados en la fase de entrenamiento está asociado a mejores resultados de las métricas *test score*, *F1-score* y AUC, aportando resultados diferentes a los obtenidos en el caso anterior.

Tabla 5: Resultados obtenidos de las métricas de evaluación para el umbral de 1000 µS/cm para los diferentes test/score.

Umbral 1000 µS/cm					
Train/Test	**Mejores modelos**	**Test score promedio**	**F1 score 0 promedio**	**F1 score 1 promedio**	**AUC promedio**
0.6/0.4	Random Forest Classifier Gradient Boosting Classifier Extra Trees Classifier	0.72	0.66	0.76	0.78
0.7/0.3	Random Forest Classifier Gradient Boosting Classifier Extra Trees Classifier	0.72	0.65	0.77	0.78
0.8/0.2	Extra Trees Classifier Gradient Boosting Classifier Support Vector Machines	0.71	0.61	0.78	0.77
0.9/0.1	Support Vector Machines K-Neightbors Logistic Regress	0.70	0.59	0.76	0.75

En el escenario que emplea un umbral de 2000 µS/cm para discriminar entre puntos positivos (aptos) y negativos (no aptos) (Tabla 6), los algoritmos que han presentado mejores resultados son *Gradient Boosting Classifier* (GBC), Support *Vector Machines* (SVC), *Random Forest Classifier* (RFC), *Quadratic Discriminant Analysis* (QDA), *Decision Tree Classifier* (DTC), *Stochastic Gradient Descent Classifier* (SGD) y *Perceptron* (PER). Destacan los algoritmos

SVC, GBC) y *RFC* con resultados óptimos en más de la mitad de los escenarios según los valores de división *train/test*. En todos los casos el *Test score* promedio logra superar el 0.76 y el AUC el 0.72, obteniendo una puntuación de hasta 0.78. Los mejores resultados para las divisiones *train/test* son los de 0.6 y 0.9, con valores de 0.79 y 0.9 respectivamente, mostrando una elevada capacidad predictiva e indicando que los extremos de división *train/test,* al igual que con el umbral de 500 µS/cm, son los mejores en las métricas generales. Sin embargo, la métrica *F1-score*, muestra valores mucho más altos para la clase positiva (1 o apto) que para la clase negativa (0 o no aptos). Los *F1-score* (1) presentan valores superiores a 0.84, mientras que los valores *F1-score* (0) no superan el 0.43, estando la mitad de los valores de división *train/test* por debajo del 0.3. Por lo tanto, se puede observar cómo ninguno de los algoritmos ha conseguido distinguir de manera precisa la clase negativa (0), las elevadas métricas generales son el resultado de que la mayoría de los puntos o muestras son positivas (aptas), lo que impide a los algoritmos diferenciar entre clases. Se puede observar cómo los mejores valores de *F1-score* varían dependiendo de si son aptos (1) o no aptos (0), los valores de *train/test* de 0.6 y 0.7 obtendrán mejores valores para *F1-score ()*, mientras que los valores de 0.8 y 0.9 obtendrán mejores valores del *F1-score* (1), indicando que contra más puntos se usen en la fase de entrenamiento mejor diferenciará o predecirá los puntos negativos, a costa de un ligero empeoramiento de predicción de los positivos.

Tabla 6: Resultados obtenidos de las métricas de evaluación para el umbral de 2000 µS/cm para los diferentes test/score.

Umbral 2000 µS/cm					
Train/Test	**Mejores modelos**	**Test score promedio**	**F1 score 0 promedio**	**F1 score 1 promedio**	**AUC promedio**
0.6/0.4	Support Vector Machines Perceptron Quadratic Discriminant Analysis	0.79	0.28	0.87	0.72
0.7/0.3	Stochastic Gradient Descent Classifier Gradient Boosting Classifier Decision Tree Classifier	0.76	0.30	0.85	0.75
0.8/0.2	Support Vector Machines Random Forest Classifier Gradient Boosting Classifier	0.76	0.43	0.84	0.78
0.9/0.1	Support Vector Machines Gradient Boosting Classifier Random Forest Classifier	0.90	0.41	0.84	0.77

El análisis de las métricas obtenidas ha permitido establecer que el mejor escenario para hacer la cartografía predictiva es el umbral intermedio, de 1000 µS/cm. Esto se debe a que, aunque el umbral de 500 µS/cm haya presentado buenos resultados respecto a las métricas test *score* y *F1-score*, este no sería de gran utilidad para elaborar la cartografía predictiva. Este umbral, con un valor de CE para diferenciar entre clases muy bajo, provoca que los algoritmos clasifiquen valores de CE más elevados (entre 500 y 1000 µS/cm) como no aptos, aunque siguen dentro de un rango de calidad bueno y apto para el consumo. Por otro lado, el umbral de 2000 µS/cm presenta valores de *test score* promedios altos, sin embargo, los valores de *F1-score* indican que no es capaz de diferenciar adecuadamente la clase negativa o no aptos para el consumo humano. Bajo este umbral, los modelos asumen que prácticamente todas las muestras presentan una CE por debajo del umbral, por lo que no es fiable realizar una cartografía predictiva útil al no ser capaz de predecir adecuadamente los valores no aptos para el consumo humano.

Dentro del escenario que emplea un umbral de 1000 µS/cm se ha optado por utilizar los resultados obtenidos a partir de la división *train/test* de 0.6. Esta elección se debe a que, a pesar de que las métricas sean similares a las obtenidas con la división *train/test* de 0.7, se ha optado por emplear la que presentaba valores *F1-score* más próximos entre clases. Estos resultados coinciden con los obtenidos por otros autores con enfoques similares como Araya et al. (2023), los cuales obtuvieron unos resultados del AUC aproximados de entre 0,79 a 0.84 empleando un enfoque similar, habiendo obtenido en este proyecto resultados de AUC de entre 0.72 a 0.8.

Por otro lado, autores como Gómez-Escalonilla *et al.* (2022) obtuvieron resultados más favorables, con valores medios de *test score* de 0.86, un *F1-score* superior a 0.82 tanto para aptos como para no aptos y un AUC medio superior a 0.9. Esta diferencia de resultados obtenidos entre ambos trabajos se puede deber a varios factores. Por un lado los algoritmos de *machine learning* dependen de la cantidad y calidad de datos disponibles, por lo que una de las posibles razones por la que obtuvieron mejores valores podría deberse a una mejor calidad de los datos introducidos en esa investigación, ya que, en este trabajo se pudo observar en el procesamiento de la base de datos de puntos de agua, como alguno de ellos, los cuales se descartaron, se situaban incorrectamente en lugares que no tenían sentido como, por ejemplo, en medio del mar, lo que podría significar que algunos puntos que si se encontraban dentro de la zona de estudio podrían llegar a tener unas coordenadas incorrectas, afectando al proceso de aprendizaje erróneamente. Esta limitación es común en bases de datos de puntos de agua de países africanos, dónde las conexiones inalámbricas en ocasiones condicionan la precisión de la georreferenciación (Gómez-Escalonilla, 2024).

Por otro lado, estos resultados se pueden ver influenciados por las variables explicativas, estando condicionados tanto por su resolución como por el tipo de variable. Las 20 variables

explicativas utilizadas en este trabajo presentan resoluciones diferentes, haciendo que algunas de ellas aporten información con una menor exactitud, generalizando un valor para un área extensa de territorio, como ocurre con la variable explicativa de la recarga. A su vez, existe la probabilidad de que no se hayan contemplado otras variables con un posible grado de importancia alto de influencia en la conductividad eléctrica, afectando de nuevo al proceso de aprendizaje, excluyendo involuntariamente información relevante. Sin embargo, tal y como se comentará más adelante, la mayor parte de las variables explicativas de mayor importancia coinciden con las obtenidas en otros trabajos.

Para la elaboración de la cartografía predictiva se han combinado las probabilidades de los 3 mejores algoritmos (*Random Forest Classifier (RFC)*, *Extra Trees Classifier* (ETC) y *Gradient Boosting Classifier* (RBC)), obteniendo nuevos valores de probabilidad medios.

4.3. Influencia de las variables explicativas

A continuación, se muestran los gráficos que representan la importancia de las variables explicativas (Figura 14) para los algoritmos con mejores resultados, el *Random Forest Classifier* (A), *Extra Trees Classifier* (B) y *Gradient Boosting Classifier* (C), para el umbral de 1000 µS/cm con un *train/test* de 0.6, el seleccionado para la realización de la cartografía predictiva.

La gráfica perteneciente al algoritmo *Random Forest Classifier* (Figura 14 A) muestra que las variables con mayor importancia para predecir la CE de las aguas subterráneas son la precipitación, la evapotranspiración real en los períodos húmedos, la distancia a la costa, la evapotranspiración real media, el nivel freático, el lecho rocoso y la topografía. La importancia que sustentan estas 7 variables combinadas es de 0,58 aproximadamente. Por otro lado, el algoritmo de *Extra Trees Classifier* indica que las variables con más peso o importancia son la evapotranspiración real en los períodos húmedos, la distancia a la costa, la precipitación, el nivel piezométrico, la topografía, la evapotranspiración real media y la distancia a las fallas (Figura 14 B).

Estas 7 variables suman aproximadamente una importancia de 0,78. Comparando estos resultados con los anteriores se observa cómo el algoritmo reparte menos el nivel de importancia entre las variables, poniendo más importancia a cinco de ellas y dando mucha menos importancia a las restantes, mientras que el anterior reparte la importancia más gradualmente, reduciendo poco a poco su nivel de variable en variable.

Por su parte, la gráfica perteneciente a la importancia de las variables explicativas para el algoritmo *Gradient Boosting Classifier* (Figura 14 C), señala que aquellas variables con mayor importancia son la distancia a la costa, la evapotranspiración real en períodos húmedos, la

precipitación, el lecho rocoso, la evapotranspiración de Thornthwaite para los períodos secos, el nivel piezométrico y la evapotranspiración real media. Presentando un valor de importancia medio de estas 7 variables de 0,68. Al igual que la anterior, se observa como la mayor importancia la sustentan 5 variables, reduciendo drásticamente en el resto, a diferencia de las dos anteriores, esta considera prácticamente irrelevantes dos de ellas, teniendo un nivel de importancia igual o muy cercano al cero, siendo estas la geología y la pendiente.

Se podría concluir que para los tres algoritmos las 7 variables con mayor importancia son más o menos las mismas exceptuando un par de ellas. Sin embargo, aunque sean las mismas cada algoritmo le da un orden de importancia diferente a cada una. Se puede observar como la precipitación, la distancia a la costa y la evapotranspiración real en periodos húmedos son siempre las 3 más importantes en los diferentes algoritmos, aunque cambien de orden.

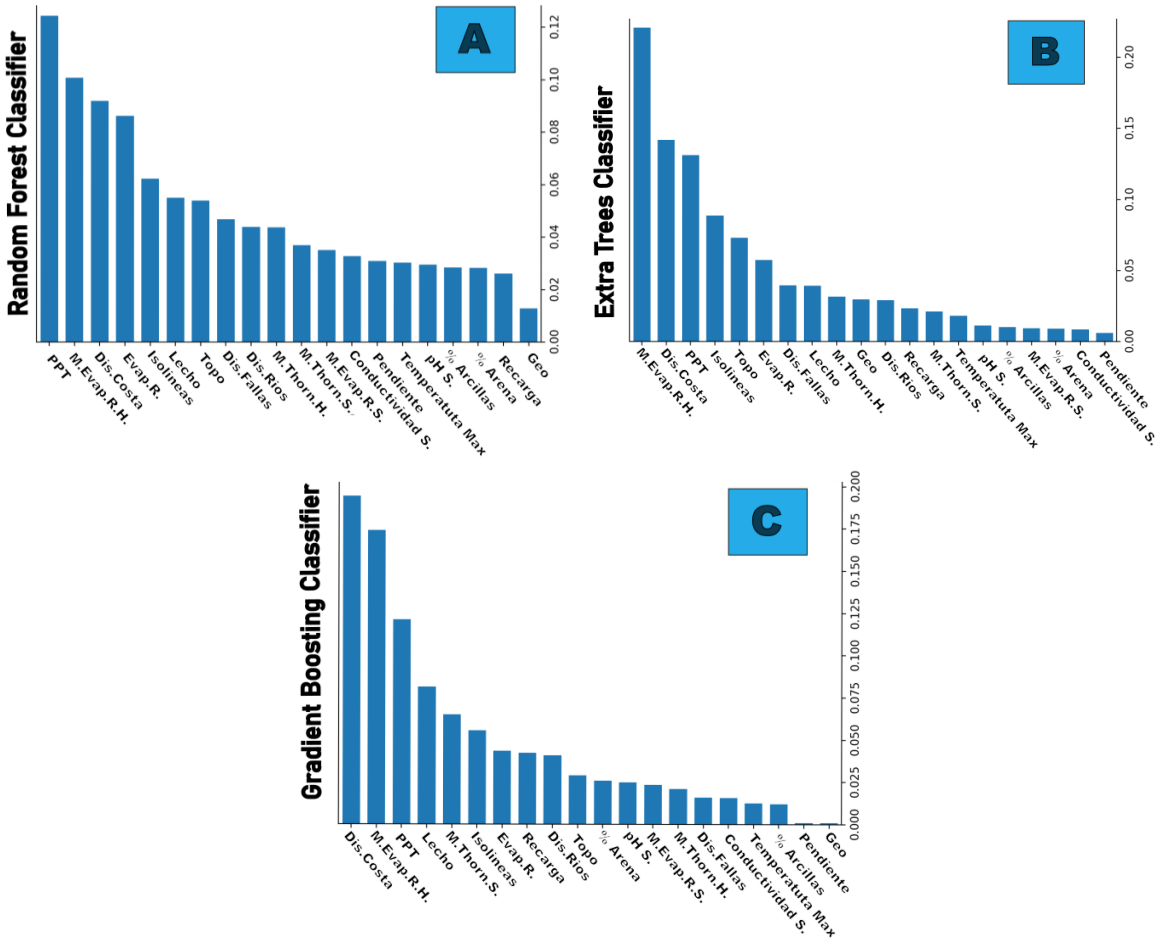

Figura 14: Importancia de las variables calculada para el umbral de 1000 µS/cm para los tres algoritmos de mejor rendimiento: A) *Random Forest Classifier*. B) *Extra Trees Classifier*. C) *Gradient Boosting Classifier.*

Estos resultados coinciden en su mayoría con otros obtenidos a través de diferentes autores como Araya et al. (2023), los cuales obtuvieron, empleando enfoques similares, la precipitación, la topografía, el lecho rocoso, la distancia a la costa y el coeficiente de *Priestley*

Taylor (relación entre evapotranspiración potencial y real) como parte de sus 7 variables más importantes. Por otro lado, para Gómez-Escalonilla (2022), en general las variables con valores con mayor peso resultaron ser la precipitación, el espesor saturado, la elevación del terreno, la geología y la densidad de drenaje, coincidiendo parcialmente con las obtenidas en este trabajo. Otros trabajos con perspectivas parecidas como Sahour et al. (2020), han obtenido variables diferentes, siendo una de ellas la transmisividad, demostrando ser para su zona de estudio la más relevante, presentando un valor de importancia mayor al 70%. Sin embargo, en este trabajo, no se contaba con información suficiente para poder elaborar una cartografía realista relacionada con este parámetro hidrogeológico, por lo que se optó por no incluirla en el procedimiento.

4.4. Cartografía predictiva

A continuación, en la Figura 15 se muestra el mapa predictivo elaborado para el umbral de 1000 µS/cm con el valor de *train/score* de 0.6. Para obtenerlo, se ha realizado el promedio de la probabilidad ofrecida por los tres mejores algoritmos *Random Forest Classifier (RFC)*, *Extra Trees Classifier* (ETC) y *Gradient Boosting Classifier* (GBC). Este mapa muestra en rojo las probabilidades bajas de que el agua subterránea sea apta, es decir, señala zonas en las que los algoritmos han encontrado un patrón de variables explicativas que les conducen a predecir una mayor probabilidad de superar el umbral establecido, en este caso 1000 µS/cm. Por el contrario, las zonas azules son aquellas en las que los algoritmos han encontrado una combinación de variables explicativas que les conducen a predecir una menor probabilidad de superar el umbral establecido, en este caso 1000 µS/cm, por lo que se consideran zonas con una conductividad eléctrica apta para el consumo humano. Se observa como las probabilidades más altas de ser aguas aptas, zonas en azul, se encuentran mayoritariamente en la zona norte y este del mapa, mientras que los más bajos se agrupan en el oeste, pegados a la costa, extendiéndose como una franja por la zona central sur. Las zonas con probabilidades más altas coinciden con elevada altura topográfica, altas precipitaciones, mayor nivel freático, mayor porcentaje de recarga, mayores tasas de evapotranspiración real y bajas temperaturas. Por el contrario, las zonas con menores probabilidades se corresponden a las zonas con valores más elevados de la temperatura, con alta tasa de evapotranspiración de Thornthwaite, con menos precipitación, menor nivel freático, y con recargas de medias a bajas. Esto indica que las precipitaciones se encuentran relacionadas con la recarga al acuífero, afectando a su vez al nivel freático, siendo un condicionante importante para la CE de este. Ya que, contra más precipitación haya más agua disponible para su infiltración, elevando el nivel freático y diluyendo los iones disueltos del agua subterránea.

Respecto a la evapotranspiración, cabe destacar que la evapotranspiración real es más elevada en las zonas en donde hay más precipitación, sin embargo, las evapotranspiraciones de Thornthwaite (evapotranspiración potencial ETP) es mucho más elevada en aquellas zonas con bajas probabilidades de ser aptas, indicando que en condiciones ideales, se evaporaría más agua de estas zonas, haciendo que incremente la posibilidad de superar el umbral de CE, dando como resultado un aumento de la conductividad eléctrica de las aguas subterráneas al saturarse de iones, esto se encuentra principalmente correlacionado con las altas temperaturas, ya que contra mayor temperatura mayor es la cantidad de agua evaporada. Por otro lado, hay una clara relación entre la elevación del terreno y la conductividad, en las zonas más elevadas hay una probabilidad mayor de que el agua sea apta en comparación con las zonas de llanuras y de costas, las cuales se encuentran a una altitud mucho menor.

Estos resultados coinciden con estudios hechos previamente en la zona como las cartografías predictivas de CRVOI (2018), las cuales indican que la mayor probabilidad de tener conductividades eléctricas elevadas en las aguas subterráneas, se encuentran al este y sur de la zona.

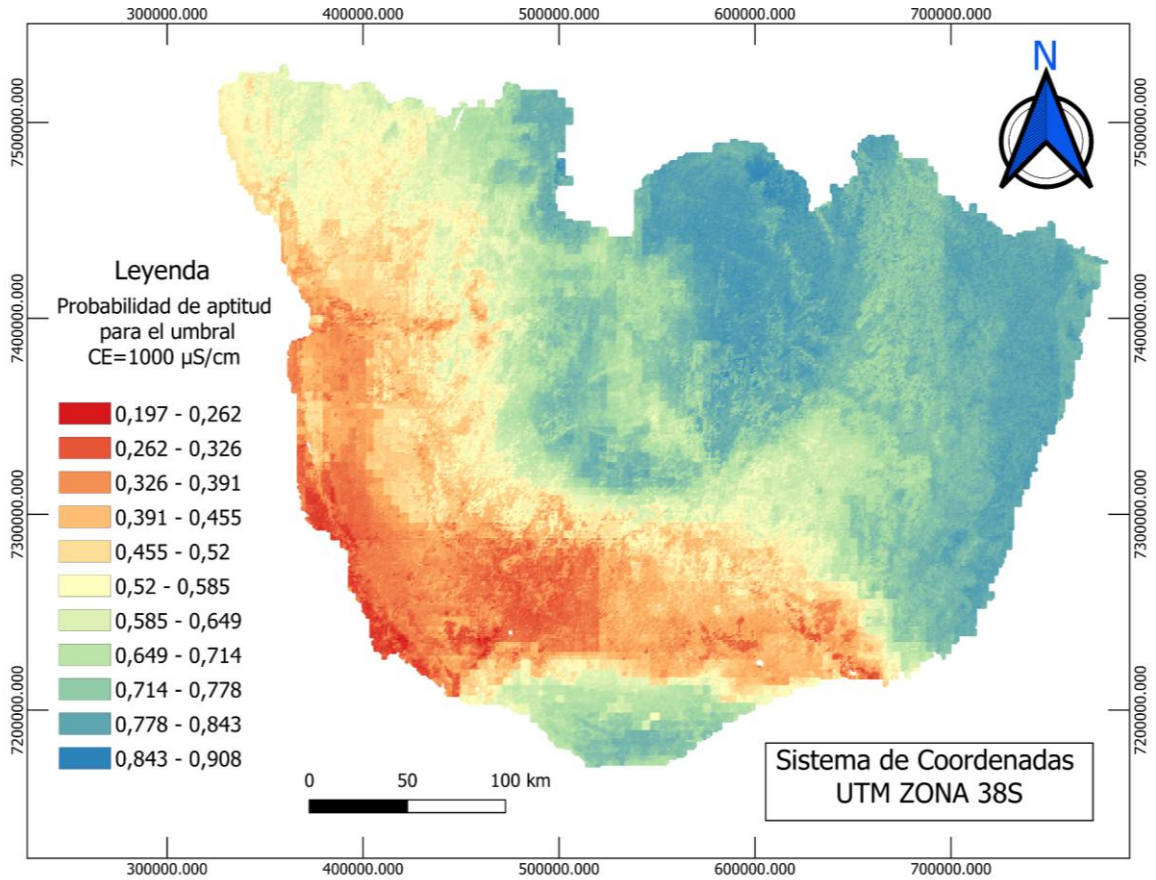

Figura 15: Mapa predictivo de la probabilidad de que el agua subterránea sea apta para el consumo, no superando el umbral de 1000 µS/cm de CE.

4.5. Estimación de población

La Figura 16 representa el mapa de densidad de población de la zona de estudio dividido por aldeas, correspondiente al año 2018. Para realizar el análisis poblacional de riesgo por consumo de agua subterránea con elevada salinidad, se ha utilizado la cartografía predictiva de la probabilidad de exceder el umbral de 1000 µS/cm, que se corresponde con la cartografía presentada en el apartado anterior. En la región, hay un total de 546 aldeas o barrios que superan la densidad de 200 habitantes por km², lo que supone menos del 17% de las aldeas totales. Entre ellas, destacan aldeas como Tanambao, Ambaro Mahazoarivo y Avaradrova. Sin embargo, toda la zona central presenta una densidad de población relativamente baja, con pequeños núcleos rurales dispersos en una amplia extensión de territorio.

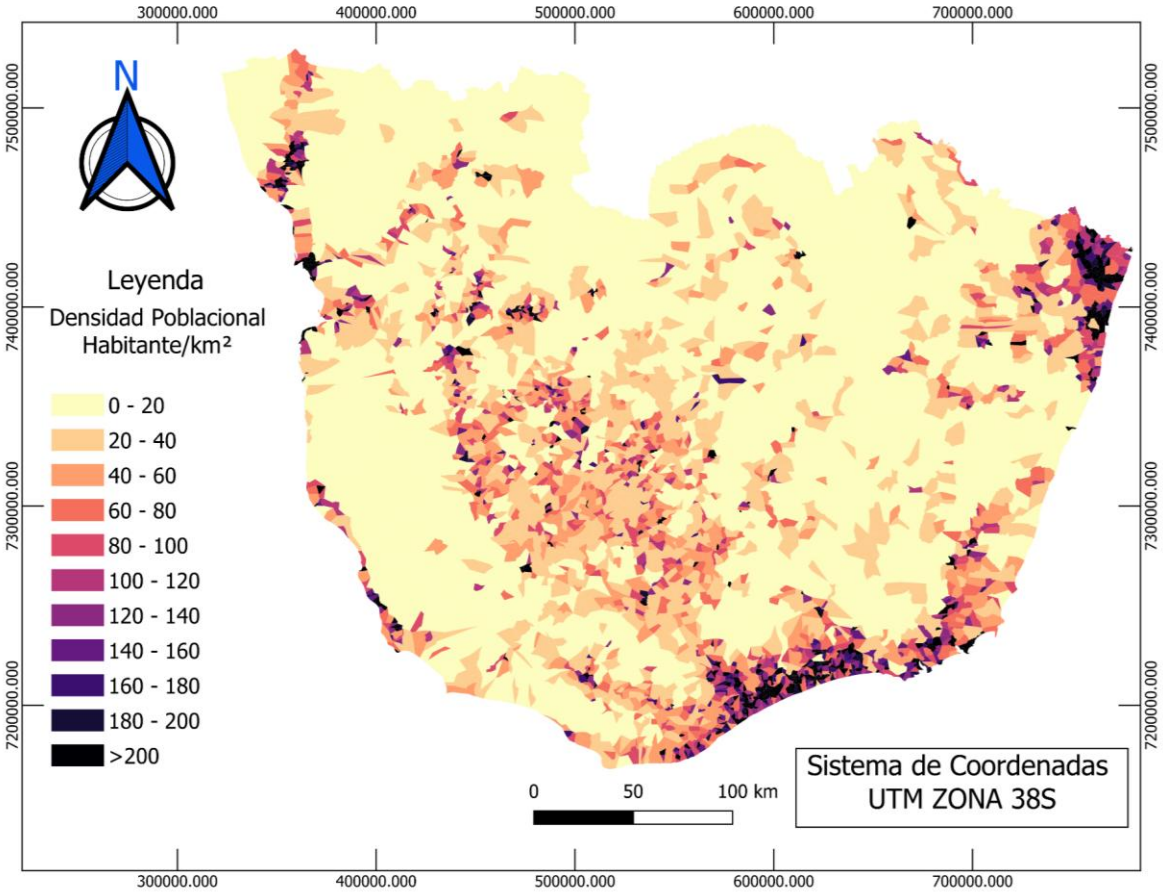

Figura 16: Mapa de densidad poblacional por cada aldea expresada en habitantes por km².

La Figura 17 muestra las diferentes aldeas y barrios categorizados por la probabilidad media de exceder el umbral de 1000 µS/cm, estando la mayoría de los núcleos más afectados al suroeste de la zona, pegados a la costa oeste.

Si comparamos los principales núcleos con mayor densidad poblacional con la probabilidad de exceder el umbral de 1000 µS/cm, se puede observar como la mayoría de estas poblaciones se encuentran en probabilidades medias y bajas a excepción de los núcleos con

mayor densidad poblacional situados al sur, los cuales se encuentran en probabilidades medias altas.

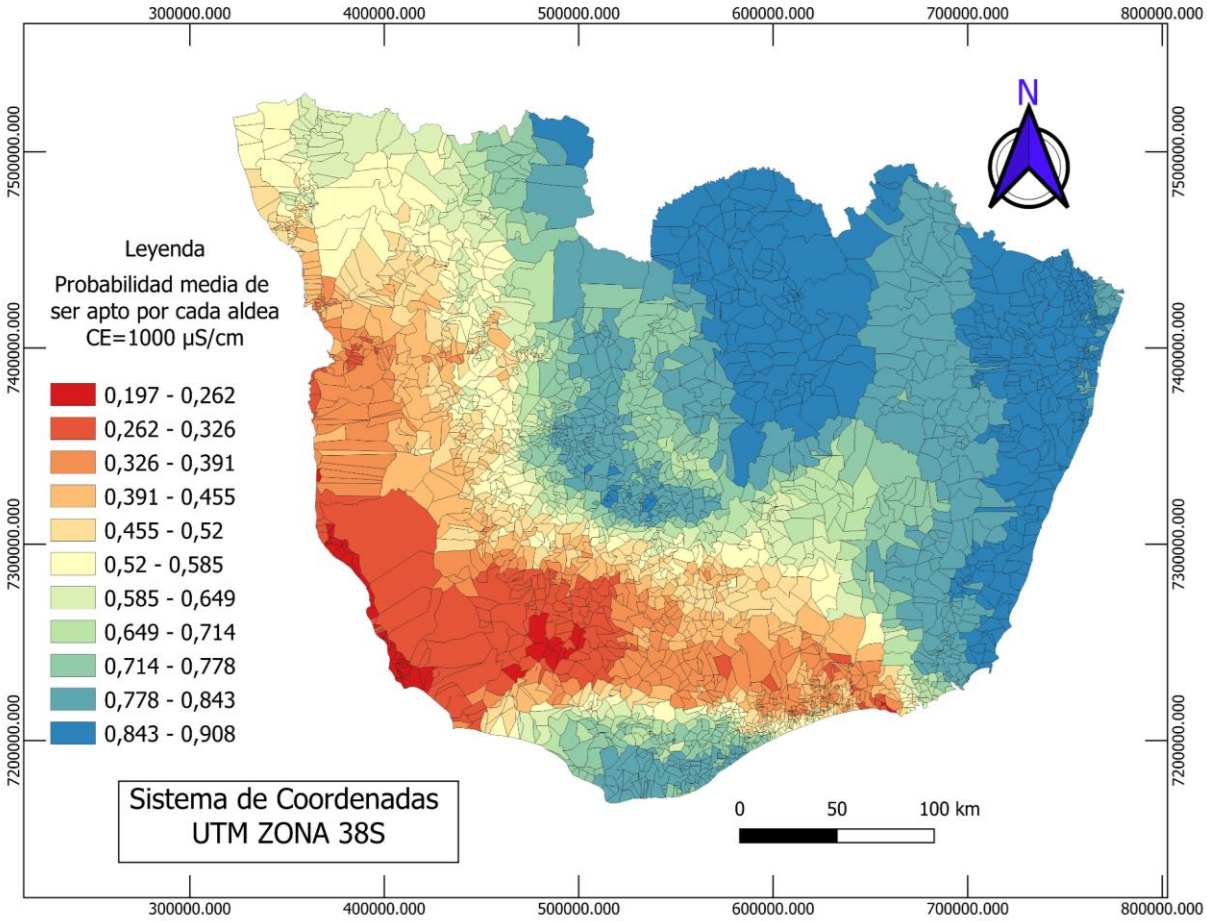

Figura 17: Mapa de probabilidad media por cada aldea de que el agua subterránea de la zona sea apta para el consumo humano para el umbral de 1000 µS/cm de CE.

Tal y como se puede observar en la Tabla 7, para el umbral de 1000 µS/cm, el 0.6% del área total presenta una probabilidad menor al 25% de ser apta para el consumo teniendo en cuenta el umbral previamente mencionado. En otras palabras, el 1.52% de la población se encuentra en zonas con alta probabilidad (>75%) de exceder este umbral. Por el contrario, el 99,4% del área y el 98.48% de población restante presentan mayores probabilidades de que el agua sea apta o no exceda el umbral de 1000 µS/cm. El 70,6% del área total tiene más probabilidades (superior al 50%) de presentar aguas aptas para el consumo humano que no superen el umbral de 1000 µS/cm de CE, lo que se corresponde con más del 67% de la población. Por último, un 44% del área total de la región estudiada presenta una probabilidad superior al 75% de que el agua sea apta y no supere el umbral de 1000 µS/cm, abarcando únicamente al 41,66% de la población total de la región. Por lo tanto, solo el 58,34% de la población se encuentra en zonas con probabilidades menores al 75% de obtener agua potable apta para

el consumo humano, ya que, la mayoría de las aldeas con altas densidades poblacionales se encuentran en zonas con bajas probabilidades de superar el umbral de 1000 µS/cm.

Tabla 7: Valores de población y area afectada para las diferentes probabilidades de que el agua sea apta para el consumo con un umbral de 1000 µS/cm

Probabilidad de ser apto CE=1000 µS/cm	Probabilidad < 25%	Probabilidad ≥ 25%	Probabilidad ≥ 50%	Probabilidad ≥ 75%
Población total	51853 (1.52%)	3352029 (98.48%)	2308832 (67.83%)	1417898 (41.66%)
Area(%) Afectada	0.60%	99.40%	70.62%	44.05%

Estos resultados en comparación con los obtenidos por Araya et al. (2023), se contemplan más optimistas, ya que, Somalia una de las zonas en las que se centra su estudio, presenta condiciones muy parecidas a la de este, sin embargo, alrededor de un 50% de la población (5 millones de personas) se encuentran expuestas a aguas con altas salinidades. En este estudio casi la mitad de la población total se encuentra en zonas con bajas probabilidades de superar el umbral de 1000 µS/cm, estando solo un 1.52% de la población afectada por probabilidades extremadamente bajas de encontrar agua apta.

5.CONCLUSIONES

En este trabajo mediante el empleo de cartografías predictivas, se han podido estimar zonas del sur de Madagascar con alta vulnerabilidad o expuestas a posibles aguas subterráneas de elevado nivel de salinidad. Los resultados obtenidos a través de los algoritmos de aprendizaje automático muestran que, por lo general, las aguas con mayor probabilidad de presentar una conductividad eléctrica más elevada son aquellas situadas en la costa oeste de la isla, coincidiendo con las altas temperaturas, bajas precipitaciones, la proximidad a la costa y zonas deprimidas topográficamente. Esto se debe, probablemente, a que al tratarse de una zona árida con escasas precipitaciones y altas temperaturas, las pequeñas cantidades de agua que se puedan llegar a infiltrar al acuífero como recarga, lo harán con altos contenidos en iones debido a las altas tasas de evapotranspiración, impidiendo que las aguas subterráneas se puedan diluir y, en último término, aumentando la concentración de iones del agua, es decir, su salinidad o conductividad eléctrica. Por otro lado, una mayor proximidad a la costa incrementará el riesgo de que las aguas subterráneas se vean afectadas por la cuña salina, la cual instruye en el interior del acuífero mezclando el agua dulce del acuífero con el agua marina y, por tanto, incrementa el contenido iónico aumentando su conductividad eléctrica.

Se ha podido observar cómo los algoritmos han tenido un mejor rendimiento con el umbral de 1000 µS/cm, esto seguramente se deba a que los puntos aptos y no aptos se encuentran cercanos a un equilibrio, siendo el umbral más próximo a tener los mismos puntos aptos y no aptos.

Los algoritmos con mejores métricas para el umbral de 1000 µS/cm han sido los algoritmos ensemble “basados en árboles” *Random Forest Classifier (RFC)*, *Extra Trees Classifier* (ETC) y *Gradient Boosting Classifier* (RBC), pudiendo diferenciar adecuadamente entre los puntos aptos de no aptos, con una tasa de acierto general por encima del 70% y superior al 75% en el caso de los puntos aptos para el consumo humano.

Las variables explicativas que han resultado ser más condicionantes para determinar la conductividad eléctrica han sido la precipitación, la evapotranspiración real, la evapotranspiración real en períodos húmedos, la distancia a la costa, la elevación topográfica y en muchos de los casos el lecho rocoso.

Los núcleos urbanos más poblados se sitúan a lo largo de la costa este, mientras que toda la zona central presenta una densidad de población baja, exceptuando algunos pequeños núcleos rurales dispersos. Por otro lado, las zonas que presentan una mayor probabilidad media de exceder el umbral establecido de 1000 µS/cm de conductividad eléctrica son aquellas situadas en el sector suroccidental de la zona de estudio, pegadas a la costa oeste

y situadas en la región de Atsimo-Andrefana. Alguna de las aldeas con elevadas densidades poblacionales, mayores a 1000 habitantes por km², y con bajas probabilidades de obtener agua dentro del umbral apto de 1000 µS/cm son Antanambao I, Andamasinny, Ambaro, Anjatocu, Antanambao III, Tanambao y Morafeno. Estas presentan más de un 67% de probabilidades de superar los 1000 µS/cm de conductividad eléctrica en su zona.

Por último, se concluye que tanto la cartografía predictiva como el análisis poblacional obtenido de este trabajo son herramientas de gran utilidad en el ámbito de gestión de recursos hídricos subterráneos, ya que, este tipo de enfoque realizado a través de algoritmos de aprendizaje automático o de inteligencia artificial, permite obtener buenos resultados predictivos en aquellas zonas en las que no se disponga ni de datos hídricos suficientes ni de la necesaria información general relacionada con la hidrogeología. Por lo tanto, podría tratarse de un pilar imprescindible para la mejoría de la calidad de vida de aquellas poblaciones con escasos recursos y alta demanda de agua, proporcionando un impacto positivo en los habitantes de países vulnerables a esta problemática.

6.BIBLIOGRAFÍA

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. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015 | Scientific Data.

Adombi, A.V.D., Chesnaux, R., & Boucher, M. A. (2021). Theory-guided machine learning applied to hydrogeology—state of the art, opportunities and future challenges. Hydrogeology Journal, 29(8), 2671-2683. https://doi.org/10.1007/s10040-021-02403-2

Akramkhanov, A., et al. (2012). The assessment of spatial distribution of soil salinity risk using neural network. *Environmental Monitoring and Assessment.*

Akter, F., Bishop, T.F.A., Vervoort, R.W. (2021). Space-time modelling of groundwater level and salinity. Science of the Total Environment, 776. Space-time modelling of groundwater level and salinity - ScienceDirect.

Alagha, J. S., et al. (2017). Integrating an artificial intelligence approach with k-means clustering to model groundwater salinity: The case of Gaza coastal aquifer (Palestine).

Al-Zoubaidy M., Monteleone M. et Bünzli, M.-A. 2023. Contribution à l'hydrogéologie du Sud de Madagascar, Base de données géoréférencées et site interactif Accès données - BushProof. Editeurs: Direction pour le Développement et la Coopération, Berne et BushProof Sàrl, A

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. https://doi.org/10.1016/j.envint.2023.107925.

Barzegar, R., et al. (2016). Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction.

BBVA. (2024). *'Machine learning': ¿qué es y cómo funciona el maestro en reconocer patrones?* BBVA. Recuperado de https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/

Centre de Recherche et de Veille sur les Maladies Emergentes dans l'Océan Indien (CRVOI). (2018). *Mapa de salinidad – Gran Sur de Madagascar* [Mapa]. Programme Solidarité Eau (PSEau). Recuperado de https://www.pseau.org/outils/biblio/resume.php?d=8204&l=es

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://arxiv.org/pdf/1106.1813.

Custodio, E., Fernández García, D., & Sánchez Vila, J. (2017). *Salinización de las aguas subterráneas en los acuíferos costeros mediterráneos e insulares españoles* (1.ª ed.). Iniciativa Digital Politècnica – Universitat Politècnica de Catalunya. https://upcommons.upc.edu/bitstream/handle/2117/111515/9788498806878.pdf.

European Space Agency (2024). *Copernicus Global Digital Elevation Model*. Distributed by OpenTopography. OpenTopography - Copernicus Global Digital Elevation Models. Accessed 2025-07-10

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

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.

Gómez-Escalonilla, V., Montero-González, E., Díaz-Alcaide, S., Martín-Loeches, M., del Rosario, M. R., & Martínez-Santos, P. (2024). A machine learning approach to site groundwater contamination monitoring wells. *Applied Water Science*, *14*(12), 250.

Hosmer, D.W., Lemeshow, S. (2000). Area under the ROC curve. Appl. Logist. Regres. 160–164.

Huang, W., & Foo, S. (2002). *Neural network modeling of salinity variation in Apalachicola River*. *Water Research, 36*(1), 356–362. Neural network modeling of salinity variation in Apalachicola River - ScienceDirect

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D.,Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L. & Tondoh, J. E. (2015). *Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions*. PloS one, 10(6), e0125814. https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/b14f76c6-2655-4aa1-92c2-d9875fde2cb3.

Jeong, H., Abbas, A., Kim, H. G., Van Hoan, H., Van Tuan, P., Long, P. T.,Lee, E., & Cho, K. H. (2024). Spatial prediction of groundwater salinity in multiple aquifers of the Mekong Delta region using explainable machine learning models. *Water Research*, *266*, 122404.

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.Dataset Record: 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)

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data 6, 15:1-15:21. https://doi.org/10.1145/2382577.2382579

MacDonald, A.M., Lark, R.M., Taylor, R.G., Abiye, T., Fallas, H.C., Favreau, G., Goni, I.B., Kebede, S., Scanlon, B.R., Sorenson, J.P.R., Tijani, M., Upton, K.A., West, C. (2020): Groundwater recharge in Africa from ground based measurements. British Geological Survey. (Dataset). Groundwater recharge in Africa from ground based measurements | NGDC Cited Data | National Geoscience Data Centre (NGDC) | Our data | British Geological Survey (BGS)

Malaxetxebarria Bengoetxea, Arene. (2024). *Salinidad de las aguas subterráneas en Mali: Predicción espacial mediante herramientas de inteligencia artificial y estimación de personas en riesgo* [Trabajo de fin de máster, Universidad Complutense de Madrid].

Martínez Santos, P., Martínez Alfaro, P. E., Montero González, E., Villarroya Gil, F., Martín-Loeches, M., Díaz Alcaide, S., & Castaño Castaño, S. (2018). Hidrogeología: principios y aplicaciones. *Hidrogeología: principios y aplicaciones*.

Martínez-Santos, P., Díaz-Alcaide, S., De la Hera-Portillo, A., & Gómez-Escalonilla, V. (2021). Mapping groundwater-dependent ecosystems by means of multi-layer supervised classification. *Journal of Hydrology*, *603*, 126873.

Martínez-Santos, P., Gómez-Escalonilla, V., Díaz-Alcaide, S., Rodríguez del Rosario, M., & Aguilera, H. (2025). A surrogate approach to model groundwater level in time and space based on tree regressors. *Appl Water Sci* 15, 206. A surrogate approach to model groundwater level in time and space based on tree regressors | Applied Water Science

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.

Gleeson, Tom & Wada, Yoshihide & Bierkens, M.F.P. & Beek, Ludovicus. (2012). Water Balance of Global Aquifers Revealed by Groundwater Footprint. Nature. 488. 197-200. 10.1038/nature11295. Groundwater footprints of aquifers that are important to agriculture... | Download Scientific Diagram

OCHA Regional Office for Southern and Eastern Africa. (2018). Madagascar - Subnational Population Statistics. Disponible en: https://data.humdata.org/dataset/cod-ps-mdg.

Oficina de Información Diplomática del Ministerio de Asuntos Exteriores, Unión Europea y Cooperación de España. (2025). Ficha país: Madagascar. Disponible en: https://www.exteriores.gob.es/documents/fichaspais/madagascar_ficha%20pais.pdf

ONU-Hábitat. (s. f.). *Agua y saneamiento en entornos urbanos*. Recuperado de Agua y Saneamiento | UN-Habitat.

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. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Rajaee, T., Ebrahimi, H., & Nourani, V. (2019). A review of the artificial intelligence methods in groundwater level modeling. Journal of Hydrology, 572, 336-351.

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*, Article 125321. https://doi.org/10.1016/j.jhydrol.2020.125321

Sahuquillo Herráiz, A. (2009). *La importancia de las aguas subterráneas*. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, 103(1), 97–114. Recuperado de https://www.rac.es/ficheros/doc/00923.pdf

Sanitation and Water for All. (2022). *Perfil de país: Madagascar* [Perfil nacional – Suministro de agua y saneamiento]. Recuperado de https://www.sanitationandwaterforall.org/sites/default/files/2022-04/SWA_Profile_Madagascar_es.pdf

Serele, C., Pérez-Hoyos, A., & Kayitakire, F. (2020). *Mapping of groundwater potential zones in the drought-prone areas of south Madagascar using geospatial techniques*. Geoscience Frontiers, 11, 1403–1413. Mapping of groundwater potential zones in the drought-prone areas of south Madagascar using geospatial techniques - ScienceDirect

Smedley, P. L. (2002). *Groundwater quality: Madagascar* (Informe técnico, 4 pp.). British Geological Survey. Recuperado de: Groundwater Quality: Madagascar

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. Machine Learning Models and Algorithms for Big Data Classification.

UNESCO. (2021, 25 de junio). *Valorar los servicios de suministro de agua y de saneamiento*. In *Informe Mundial sobre el Desarrollo de los Recursos Hídricos de 2021: El valor del agua*. Recuperado de https://www.unesco.org/reports/wwdr/2021/es/valorar-los-servicios-de-suministro-de-agua-y-de-saniamento

UNESCO World Water Assessment Programme. (2022). *Informe Mundial de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos 2022: Aguas subterráneas – hacer visible el recurso invisible* (266 p.; ISBN 9789233001930). AGUAS SUBTERRÁNEAS.

UNICEF. (2018). *Uso de SIG y teledetección para acceder al agua en las zonas propensas a la sequía de Etiopía y Madagascar*. Fondo de las Naciones Unidas para la Infancia (UNICEF). https://www.unicef.org/ethiopia/media/171/file

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.

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.

United States Central Intelligence Agency. (1981). *Madagascar (Shaded Relief)* [Mapa]. Perry–Castañeda Library Map Collection, University of Texas at Austin. Recuperado de https://maps.lib.utexas.edu/maps/madagascar.html

Upton, K., Ó Dochartaigh, B., Monteleone, M., & Bellwood-Howard, I. (2018). *Africa Groundwater Atlas: Hydrogeology of Madagascar*. British Geological Survey. Recuperado de https://earthwise.bgs.ac.uk/index.php/Hydrogeology_of_Madagascar

WHO, A global overview on national regulations and standards for drinking water quality, second edition. Geneva: World Health Organizatión; 2021. Licence:CC BY-NC-SA 3.0 IGO. https://iris.who.int/bitstream/handle/10665/350981/9789240023642-eng.pdf?sequence=1 .

WorldAtlas. (2025). *The Mangoky River is the longest river in Madagascar, flowing west from the Central Highlands into the Mozambique Channel*. En *Longest Rivers on Madagascar*. Recuperado de https://www.worldatlas.com/articles/longest-rivers-on-madagascar.html

World Bank. (2025). *Population, total – Madagascar* (SP.POP.TOTL). Retrieved from datos.bancomundial.org

Zheng, A. & Casari, A. (2018). Feature Engineering for Machine Learning. O'Reilly Media, Inc. ISBN: 9781491953242. https://www.repath.in/gallery/feature_engineering_for_machine_learning.pdf