1. Laguna de Fuente de Piedra [2].

Predicción sobre el entorno medioambiental en la Laguna de Fuente de Piedra

Miguel Ángel Pérez García
Máster en Ciencia de Datos
Área 5

Nombre Consultor/a
Carlos Luis Sanchez Bocanegra

Fecha de entrega: 2022 05 29

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción sobre el entorno medioambiental en la Laguna de Fuente de Piedra*
Nombre del autor:	*Miguel Ángel Pérez García*
Nombre del consultor/a:	*Carlos Luis Sánchez Bocanegra*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	05/2022
Titulación:	*Máster en Ciencia de datos*
Área del Trabajo Final:	*TFM Área 5*
Idioma del trabajo:	*Castellano*
Palabras clave	*Análisis, predicción, medioambiente.* *Máximo 3 palabras clave, validadas por el director del trabajo (dadas por los estudiantes o en base a listados, tesauros, etc.)*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El agua es un recurso valioso, tanto es así que las aves dedican parte de sus labores a criar a sus crías en las lagunas de nuestro medio ambiente, humedales u otras áreas.

La laguna de Fuente de Piedra de Antequera, que [2] "alberga la mayor colonia de flamencos comunes, es la mayor de la Península Ibérica y la segunda en importancia de Europa". Uno de los principales desafíos es poder predecir el nivel de agua que puede tener, a partir de un cierto nivel podremos observar si habrá anidamiento de estos flamencos. Además disponemos de datasets con información de los pozos de la cuenca, que podremos analizar.

Para ello se han posicionado distintos dispositivos de medida, una estación meteorológica y otras tomas manuales. La metodología que se utiliza es ágil, donde se creará un prototipo que se irá mejorando.

A partir de modelos de aprendizaje automático se predice el nivel del agua como variable continua, aplicando técnicas de regresión. Obteniendo como resultado un 60% de precisión con el algoritmo Random Forest Regressor.

Abstract (in English, 250 words or less):

Water is a valuable resource, so much so that birds dedicate part of their work to raising their young in the surrounding lagoons, wetlands or other areas.

The Fuente de Piedra lagoon in Antequera, which [2] "houses the largest colony of common flamingos, is the largest in the Iberian Peninsula and the second in Europe". One of the main challenges is to be able to predict the level of water that it may have, from a certain level we will be able to observe if there will be nesting of these flamingos. We also have data sets with information from the wells in the basin, which we can analyze.

For this, different measuring devices, a weather station and other manual plugs have been placed. The methodology used is agile, where a prototype will be created that will be improved.

From machine learning models, the water level is predicted as a continuous variable, applying regression techniques. Obtaining as a result a 60% accuracy with the Random Forest Regressor algorithm.

Agradecimientos

A mi familia, por su constancia y ánimos.
A África Lupión y Miguel Ángel Martín por los datos de la Junta de Andalucía y por contarnos su conocimiento sobre el medio ambiente y la Laguna. En todo momento han colaborado con las dudas.
A mi tutor del proyecto, Carlos Luís Sánchez, por su propuesta y su conocimiento.

Índice

1. Introducción

1.1 Contexto y justificación del Trabajo

La Laguna de Fuente de Piedra, el humedal interior más grande que tenemos en Andalucía, es un bien medioambiental que no se ha examinado ni analizado en su totalidad. Se capturan los datos diariamente aunque no se explotan.

El número de aves que nos visitan anualmente y que crian en este humedal es indeterminado, depende de factores medioambientales, como por ejemplo si llueve o si se seca a unos niveles el humedal.

La laguna es visitada por [2] "la mayor colonia de flamencos comunes de la Península Ibérica y la segunda en importancia de Europa. Ha llegado a registrar **20.000 parejas reproductoras de flamencos**, siendo la primavera, la mejor época para observarlos, especialmente a primeras horas de la mañana. El anillamiento de estas aves congrega cada año a numerosos participantes que colaboran en esta actividad científica."

Es por ello, que se podrían aplicar técnicas de aprendizaje automático para elaborar modelos que nos ayuden a predecir el nivel de agua de la laguna, a partir de cierto nivel podremos observar si habrá anidamiento de flamencos o también predecir el número de flamencos que nos visitarán este año.

1.2 Objetivos del Trabajo

Objetivo Principal

- OP. Estimar el nivel de la laguna para saber si habrá anidamiento de aves.

Objetivos secundarios:

- S1. Predecir el nivel freático en la laguna.
- S2. Predecir el número de especies.
- S3. Predicción de la meteorología.

1.3 Enfoque y método seguido

La Junta de Andalucía tiene algunas fuentes de datos, como las estaciones meteorológicas, limnígrafo, pozos y censo de aves. Todos estos datos, excepto los del limnígrafo son capturados manualmente.

La estrategia que se propone es desarrollar un modelo para cumplir con los objetivos, siguiendo una metodología de desarrollo ágil, por iteraciones. Teniendo como bloques de trabajo; investigación sobre el dominio, captura de datos, la limpieza de datos, que en nuestro caso tiene una gran envergadura, el procesamiento de datos (integración de datos, PCA, análisis, etc.), implementación del modelo y pruebas.

1.4 Planificación del Trabajo

1. Planificación. Herramienta online para crear diagramas [1].

1.5 Breve sumario de productos obtenidos

Se definen una serie de modelos obtenidos mediante algoritmos de aprendizaje automático para predecir el nivel de La Laguna de Fuente de Piedra. Durante el proceso de obtención de los modelos; se integran los datos de la fuente, se tratan y se analizan, además de aplicar aprendizaje automático.

1.6 Breve descripción de los capítulos de la memoria

En el resto de capítulos, se hablará de la metodología del trabajo, cuáles son los datos, cómo son los datos y la implementación de los modelos. También se habla de las conclusiones y las líneas de futuro, entre otras, la mejora de las fuentes de datos para una mayor calidad.

2. Estado Del Arte

Se aplicarán técnicas a las fuentes propias de la Junta de Andalucía que posteriormente analizaremos y definiremos. En esta sección se detalla lo que se ha hecho por parte de la comunidad que tenga como objetivo nuestro proyecto o proyectos similares.

Así por ejemplo, un equipo de investigación este año ha publicado un estudio, donde se aplican técnicas de aprendizaje automático como herramienta sostenible para la predicción de oxígeno disuelto en el embalse de Feitsui, Taiwán[3]. Se trabaja con datos históricos sobre el embalse de Feitsui que la administración posee, similar al objetivo S1 descrito anteriormente. Además, se utiliza Deep Learning para predecir, con lo cual es interesante valorar esta opción para nuestro proyecto.

Hace seis años atrás también se estudió algo parecido, aplicando redes neuronales artificiales para la predicción de la calidad del agua en el océano[5]. Aunque nuestros objetivos van más en la línea de predecir el nivel freático en función de variables ambientales, algo parecido analiza Olivier Lejeune (2020) con una red neuronal en el río Rin[5].

Si nos enfocamos en los estudios sobre el anidamiento de aves, son varios los que utilizan técnicas de aprendizaje automático. Varios científicos de Asia han analizado en las tierras altas la reproducción para la Grulla de cuello negro con Species Distribution Models (Modelos de Distribución de especies) [7]. Además existen modelos que utilizan CNN (Redes Neuronales Convolucionales) para la detección de aves sobre el medio ambiente[8].

Sin embargo, son muchos los investigadores y estudios que desde hace muchos años dan cobertura al análisis y predicción del clima, así que para el S3 es fácil encontrar trabajos de científicos que estudian el tiempo con técnicas de aprendizaje automático. Como lo explica este artículo del IEEE, “El Centro Nacional de Investigación Atmosférica (NCAR) tiene una larga historia de aplicación del aprendizaje automático a los desafíos del pronóstico del tiempo” [9]. Se cita un estudio con datos del Sistema de Pronóstico Global (GFS) donde se pronostica el clima en la Cuenca hidrográfica en Canadá con tres algoritmos diferentes (red neuronal bayesiana (BNN), regresión vectorial de soporte (SVR) y proceso gaussiano (GP) [10], para nuestro trabajo puede ser útil implementar también más de un algoritmo de aprendizaje y comprobar cuál genera mejores resultados.

3. Metodología

3.1 Análisis

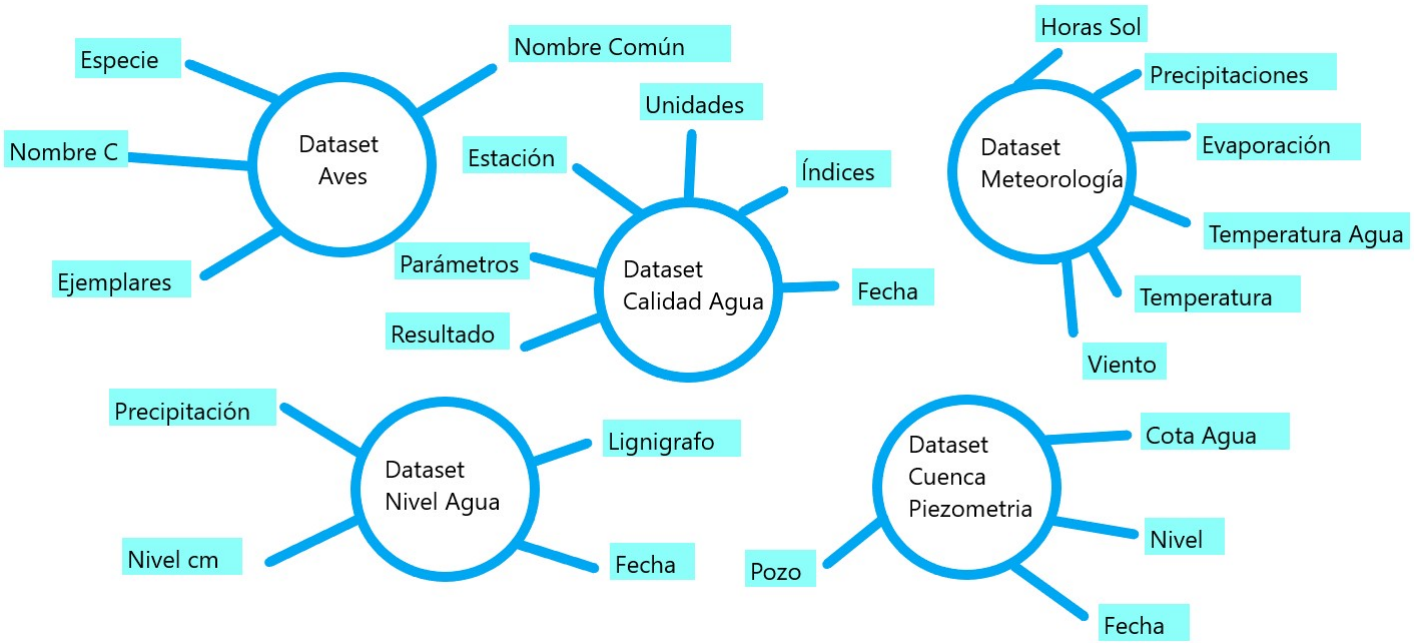
3.1.1 Análisis de fuentes

La Junta de Andalucía nos proporciona un conjunto de archivos en formato XLS y ODS. Este formato hace el trabajo más complejo (formatos más manejables son por ejemplo CSV o JSON).

Propiedades:

- Espacio en disco: 13 MB.
- Número de ficheros: 235.
- Formato: XLS y ODS.

A continuación, en la imagen 2 podemos apreciar cuáles son los dominios y atributos más interesantes que podemos explorar:



2. Diagrama de datasets. Elaboración propia.

La fuente de datos de la que partimos es pobre en cuanto a calidad del dato. Hay ficheros que tienen duplicados, las tablas son irregulares como en las figuras 3 y 4, no están estructuradas muchas de ellas no tienen formato tabular, algunos registros tienen nulos como en la figura 5, etc.

Definimos como requisito predecir el nivel del agua de la laguna, en la figura 6 se ilustran algunos conceptos que utilizaremos como atributos en nuestro dataset.

3.1.2 Ejemplos de fuentes

1	INSTITUTO NACIONAL DE METEOROLOGIA											
2												
3												
4	OBSERVATORIO DE				Cerro del Palo (Fuente de Piedra)				PROVINCIA		Málaga	
5	EVAPORACION DIARIA EN TANQUE TIPO											
6	ALTURA DE LA BOCA DEL PLUVIOMETRO SOBRE				31 cm.				INDICATIVO HID.			
7	DISTANCIA DEL PLUVIOMETRO AL CENTRO DEL				75 cm.				HORA OBS.		8 h.s.	
8	ALTURA DEL ANEMOMETRO SOBRE EL SUELO				78 cm.				MES		FEBRERO	
9	DISTANCIA DEL ANEMOMETRO AL CENTRO DEL				70 cm.				AÑO		1996	
10												
11												
12	1	2	3	4	5	6	7	8	9	10	11	
13	FECHA	RECIPITAC	NIVEL	NIVEL	NIVEL	EVAPORAC	RECORRIDO VIENT		TEMPERATURA A		CONTINGECIA	
14		24 HOR.		DE	LEIDO	24 HORAS	KM LEIDO	EN 24 H.	MAXIMA	MINIMA	OCURRIDA	
15	LECTURA	NTERIORE	NTERIO	EFERENCI	(mm.)	NTERIORES		ANTERIORE	24 HORAS ANTERIORES			
16	2	2,5	67,55	65,05	67,55	2,5			15	6	LLUVIA	
17	3	0	67,55	67,55	68,9	1,35			12	3		
18	4	0	68,9	68,9	69,3	0,4			13	5		
19	5	0	69,3	69,3	70,4	1,1			10	4		
20	6	3,5	70,4	66,9	66,6	-0,3			15	3	LLUVIA	

3. Datos en fichero XLS. No existe un formato tabular homogéneo. Con valores nulos.

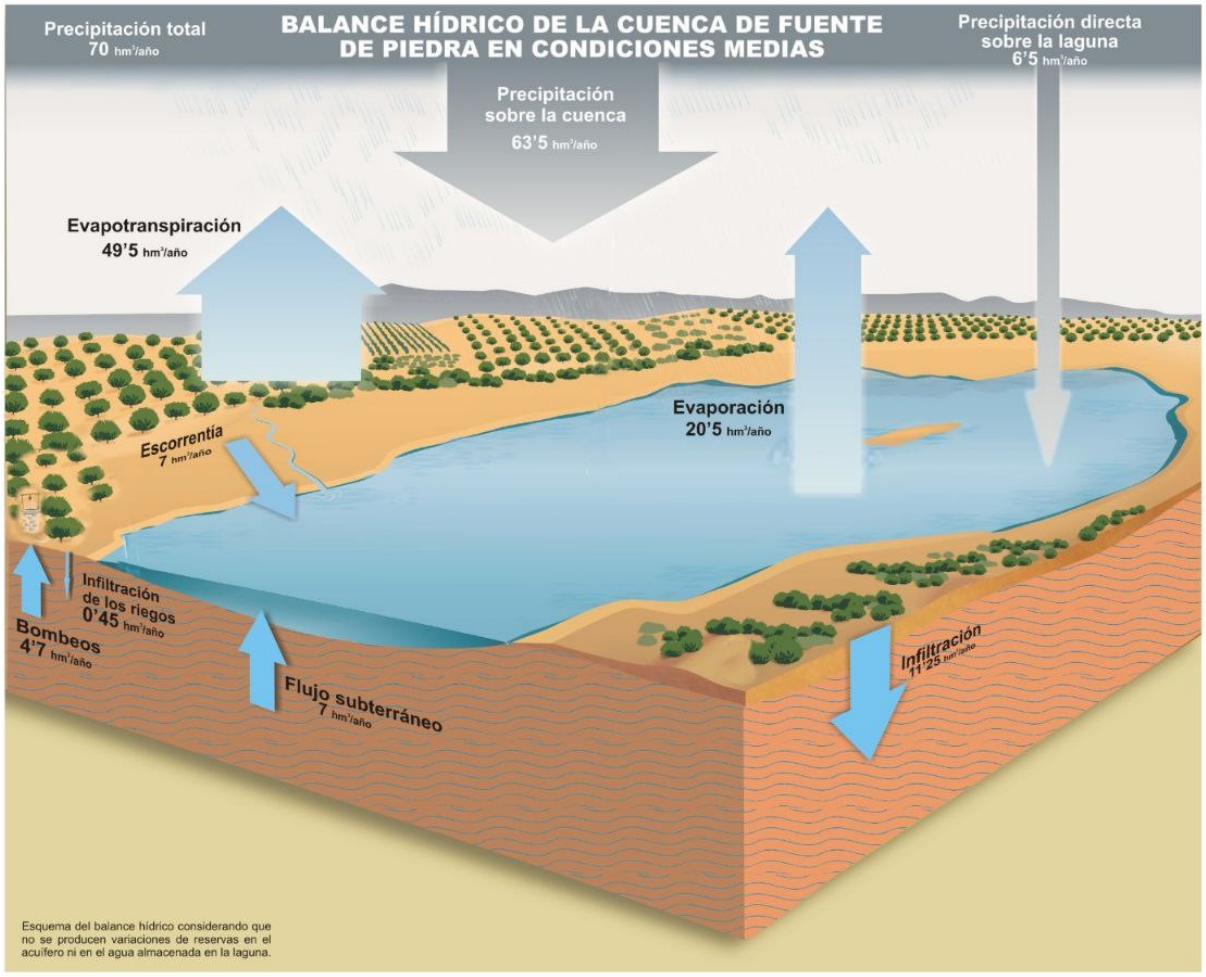
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1															FECHA	NIVEL	PRECIPITAC.	EVAPORA	NIVEL	LIMNIGRAFO	
2	DIA	OCT	NOV	DIC	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP		1-10-89				-16	2,56	
3	1	2,6	2,3	1,8		1,1	1,1		1,2	1,3	1,5		1,9		2-10-89				-17	2,57	
4	2	2,6	2,3	1,8		1,1	1,1		1,2	1,3	1,5		2		3-10-89				-17	2,57	
5	3	2,6	2,4	1,8		1,1	1,1	1,1	1,2	1,3			2		4-10-89				-17	2,57	
6	4	2,6	2,4	1,8		1,1	1,1	1,1	1,2	1,3			2		5-10-89				-17	2,57	
7	5	2,6	2,4	1,7		1,1	1,1	1,1	1,2	1,3			2		6-10-89				-17	2,57	
8	6	2,6	2,4	1,7		1,1	1,1	1,1	1,2	1,3			2		7-10-89				-17	2,57	
9	7	2,6	2,4	1,7		1,1	1,1	1,1	1,2	1,4			2		8-10-89				-17	2,57	
10	8	2,6	2,4	1,7		1,1	1,1	1,1	1,2	1,4			2		9-10-89				-17	2,57	
11	9	2,6	2,4	1,7		1,1	1,1	1,1	1,2	1,4			2		######				-17	2,57	
12	10	2,6	2,4	1,6		1,1		1,1	1,2	1,4			2		######				-18	2,58	
13	11	2,6	2,4	1,6		1,1		1,1	1,2	1,4			2		######				-18	2,58	
14	12	2,6	2,4	1,6		1,1		1,1	1,2	1,4			2		######				-18	2,58	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	DIA	octubre-02	noviembre-02	########	########	febrero-03	marzo-03	#####	mayo-03	junio-03	#####	agosto-03	septiembre-03		FECHA	NIVEL	PRECIPITAC.	EVAPORA	NIVEL	LIMNIGRAFO	
2	1	2,62	2,52	2,1	2,05	2,02	2	1,99	2,07	2,25	2,45	2,56	2,64		1-10-02				-22	2,62	
3	2	2,54	2,53	2,1	2,05	2,02	2	2	2,07	2,25	2,45	2,56	2,64		2-10-02				-14	2,54	
4	3	2,44	2,53	2,11	2,05	2,03	2	2	2,09	2,25	2,46	2,57	2,64		3-10-02				-4	2,44	
5	4	2,43	2,54	2,11	2,05	2,03	2	2,01	2,1	2,26	2,46	2,57	2,64		4-10-02				-3	2,43	
6	5	2,43	2,54	2,11	2,05	2,03	2,01	2,03	2,1	2,28	2,48	2,58	2,65		5-10-02				-3	2,43	
7	6	2,44	2,55	2,11	2,05	2,03	2,01	2,02	2,1	2,29	2,48	2,58	2,65		6-10-02				-4	2,44	
8	7	2,45	2,56	2,13	2,05	2,03	2,01	2,02	2,1	2,29	2,48	2,58	2,66		7-10-02				-5	2,45	
9	8	2,46	2,56	2,13	2,04	2,03	2,01	2,04	2,1	2,29	2,49	2,58	2,66		8-10-02				-6	2,46	
10	9	2,41	2,57	2,12	2,04	2,03	2,02	2,04	2,1	2,31	2,49	2,59	2,67		9-10-02				-1	2,41	
11	10	2,39	2,57	2,1	2,03	2,03	2,03	2,04	2,11	2,31	2,49	2,59	2,67		######				1	2,39	

4. Pestañas del año 90 y 02 con distinto formato para el mismo fichero XLS.

Ob_1642-2-071	F_1642-2-092	N_1642-2-092	Ob_1642-2-092	F_1642-2-097	N_1642-2-097	Obv_1642-2-097	F_1642-2-151	N_1642-2-151	Obv_1642-2-151	F_1642-2-236	N_1642-2-236	Obv_1642-2-236	F_1642-2-255	N_1642-2-255	Obv_1642-2-255	F_1642-2-256	N_1642-2-256	Obv_1642-2-256	F_1642-2-282	N_1642-2-
436,79	30/06/1991	4,9	410,38	12/11/1998	6,28	403,72	30/03/1992	10,92	437,08	08/09/1998	5,45	414,55	28/07/1994	4,04	415,96	03/01/1996	11,06	459,95	13/11/1998	
436,38	31/07/1991	4,86	410,43	09/12/1998	6,24	403,76	28/04/1992	11,05	436,95	12/11/1998	5,52	414,48	29/09/1994	4,19	415,81	16/10/1996	8	463,01	10/12/1998	
436,16	31/08/1991	5,6	409,68	25/01/1999	6,13	403,87	27/05/1992	9,05	438,95	09/12/1998	5,65	414,35	12/01/1995	4,28	415,72	16/10/1997	8	463,01	25/01/1999	
436,25	26/09/1991	5,62	409,66	03/05/1999	6,29	403,71	13/08/1992	9,95	438,05	25/01/1999	5,64	414,36	16/02/1995	4,25	415,75	13/01/1998	6,21	464,8	03/05/1999	
436,54	29/10/1991	5	410,28	05/10/1999	8,14	401,86	26/09/1992	10	438	03/05/1999	6,26	413,74	17/03/1995	4,18	415,82	14/02/1998	5,76	465,25	04/10/1999	
438,51	28/11/1991	4,85	410,43	18/05/2000	2,51	407,49	21/01/1994	10,7	437,3	05/10/1999	8,4	411,6	17/04/1995	4,15	415,85	20/02/1998	6,1	464,91	18/05/2000	
438,54	01/01/1992	4,7	410,58	30/10/2000	7,06	402,94	22/02/1994	10,7	437,3	18/05/2000	7,96	412,04	23/05/1995	4,14	415,86	18/03/1998	5,6	465,41	30/10/2000	
438,04	27/02/1992	4,78	410,5	23/01/2001	5,88	404,12	08/04/1994	11	437	30/10/2000	9,76	410,24	28/06/1995	4,15	415,85	22/04/1998	6,32	464,69	23/01/2001	
437,73	30/03/1992	5,15	410,13	21/09/2001	7,02	402,98	19/05/1994	11,14	436,86	23/01/2001	9,33	410,67	24/07/1995	4,24	415,76	29/05/1998	6,32	464,69	21/09/2001	
437,53	28/04/1992	5,3	409,98	06/10/2001	5,59	404,41	26/06/1994	12,1	435,9	21/09/2001	11,06	408,94	01/08/1995	4,19	415,81	02/06/1998	6,11	464,9	06/10/2001	
437,14	27/05/1992	4,74	410,54	07/11/2001	5,59	404,41	30/07/1994	11,94	436,06	06/10/2001	10,05	409,95	31/08/1995	4,3	415,7	30/06/1998	5,95	465,06	07/11/2001	
436,59	12/08/1992	5,6	409,68	08/02/2002	5,3	404,7	05/08/1994	12,15	435,85	07/11/2001	10,12	409,88	28/09/1995	4,37	415,63	30/07/1998	6,23	464,78	08/02/2002	
436,07	26/09/1992	5,6	409,68	08/03/2002	5,26	404,74	29/09/1994	11,63	436,37	08/02/2002	9,9	410,1	31/10/1995	4,38	415,62	08/09/1998	6,52	464,49	08/03/2002	
436,03	19/01/1994	5,21	410,07	08/05/2002	5,74	404,26	12/01/1995	11,7	436,3	08/03/2002	9,88	410,12	28/11/1995	4,43	415,57	16/11/1998	6,34	464,67	08/05/2002	
436,03	23/02/1994	5,06	410,22	08/06/2002	6,61	403,39	17/02/1995	11,68	436,32	08/05/2002	9,96	410,04	03/01/1996	4,23	415,77	11/12/1998	6,44	464,57	08/06/2002	
435,94	08/04/1994	4,98	410,3	08/09/2002	7,21	402,79	15/03/1995	11,86	436,14	08/06/2002	10,41	409,59	12/02/1996	1,6	418,4	25/01/1999	6,49	464,52	08/09/2002	
436,03	19/05/1994	6,27	409,91	01/11/2002	6,9	403,1	14/04/1995	12,33	435,67	01/11/2002	12,72	407,28	14/10/1997	2,6	417,4	03/05/1999	7,03	463,98	01/11/2002	
436,16	23/06/1994	6,43	408,85	01/02/2003	5,84	404,16	23/05/1995	11,55	436,45	01/02/2003	11,45	408,55	01/12/1997	2,13	417,87	05/10/1999	8,29	462,72	01/02/2003	
436,08	30/07/1994	6,54	408,74	01/05/2003	5,47	404,53	28/06/1995			01/05/2003	11,17	408,83	12/02/1998	1,59	418,41	18/05/2000	7,84	463,17	01/05/2003	
435,54	01/08/1994	6,68	408,6	01/06/2003	6,4	403,6	25/07/1995			01/06/2003	11,4	408,6	19/02/1998	1,78	418,22	30/10/2000	8,91	462,1	01/06/2003	
435,04	06/09/1994	6,67	408,61	01/09/2003	7,35	402,65	31/08/1995			01/09/2003	13,36	406,64	18/03/1998	2,07	417,93	23/01/2001	8,04	462,97	01/09/2003	
437,34	11/01/1995	6,27	409,01	01/10/2003	6,91	403,09	28/09/1995			01/10/2003	12,83	407,17	22/04/1998	2,23	417,77	21/09/2001	8,35	462,66	01/10/2003	
437,13	17/02/1995	6,07	409,21	01/01/2004	6,6	403,4	30/10/1995			01/01/2004	10,3	409,7	20/05/1998	2,26	417,74	06/10/2001	7,64	463,37	01/01/2004	
437,61	15/03/1995	6,15	409,13	01/03/2004	5,07	404,93	30/11/1995			01/03/2004	10,23	409,77	02/06/1998	2,35	417,65	07/11/2001	8,2	462,81	01/03/2004	
437,6	17/04/1995	6,5	408,78	01/04/2004	5,14	404,86	03/01/1996	12,4	435,6	01/04/2004	9,24	410,76	30/06/1998	2,64	417,36	08/02/2002	7,13	463,88	01/04/2004	
437,32	24/05/1995	6,8	408,48	01/06/2004	5,56	404,44	01/03/1996	11,88	436,12	01/06/2004	9,3	410,7	29/07/1998	2,87	417,13	08/03/2002	7,4	463,61	01/06/2004	
436,61	27/06/1995	6,95	408,33	01/07/2004	5,93	404,07	01/04/1996	12	436	01/07/2004	10,15	409,85	07/09/1998	3,04	416,96	08/05/2002	7,62	463,39	01/11/2004	
436,19	24/07/1995	7,06	408,22	01/09/2004	6,51	403,49	01/05/1996	10,79	437,21	01/09/2004	10,57	409,43	12/11/1998	3,06	416,94	08/06/2002	8,41	462,6	01/12/2004	
436,07	31/08/1995	7	408,28	01/11/2004	6,13	403,87	01/06/1996	13,8	434,2	01/10/2004			09/12/1998	3,07	416,93	01/11/2002	8,2	462,81	01/01/2005	
436,65	28/09/1995	7,33	407,95	01/12/2004	5,88	404,12	01/07/1996	13,8	434,2	01/11/2004	10,85	409,15	25/01/1999	3,02	416,98	01/02/2003	7,96	463,05	01/02/2005	
437,82	31/10/1995	7,4	407,88	01/01/2005	5,77	404,23	01/08/1996	12,15	435,85	01/12/2004	10,77	409,23	03/05/1999	3,02	416,98	01/05/2003	7,97	463,04	01/03/2005	
437,27	28/11/1995	7,35	407,93	01/03/2005	5,33	404,67	01/09/1996	11,58	436,42	01/01/2005	10,73	409,27	05/10/1999	3,94	416,06	01/09/2003	8,02	462,99	01/04/2005	
436,56	03/01/1996	5,74	409,54				01/10/1996	11,4	436,6	01/02/2005	10,71	409,29	18/05/2000	2,98	417,02	01/10/2003	8,68	462,33	01/05/2005	
437,14	09/02/1996	3,13	412,15				01/11/1996	11,03	436,97	01/03/2005			30/10/2000	3,74	416,26	01/01/2004	8,42	462,59	01/06/2005	
436,55	15/10/1997	3,77	411,51				01/02/1997	8,65	439,35	01/04/2005	11,13	408,87	23/01/2001	2,96	417,04	01/03/2004	7,38	463,63	01/07/2005	
437,01	28/11/1997	3,18	412,1				01/03/1997	7,8	440,2	01/05/2005	11,39	408,61	21/09/2001	3,64	416,36	01/04/2004	7,6	463,41	30/08/2005	
436,74	12/01/1998	1,75	413,53				01/04/1997	7,38	440,62	01/06/2005	11,65	408,35	06/10/2001	3,11	416,89	01/06/2004	7,58	463,43	01/09/2005	
436,92	12/02/1998	1,35	413,93				01/05/1997	6,94	441,06	01/07/2005	12,37	407,63	07/11/2001	2,93	417,07	01/02/2005	8,53	462,48	01/10/2005	
436,42	19/02/1998	1,5	413,78				01/06/1997	7	441	30/08/2005	12,63	407,37	08/02/2002	2,75	417,25	01/03/2005	8,4	462,61	01/11/2005	
435,66	18/03/1998	2,04	413,24				01/07/1997	6,68	441,32	01/09/2005	12,95	407,05	08/03/2002	2,91	417,09	01/04/2005	8,85	462,16	01/12/2005	

5. Datos sin completar en el fichero de piezometría.

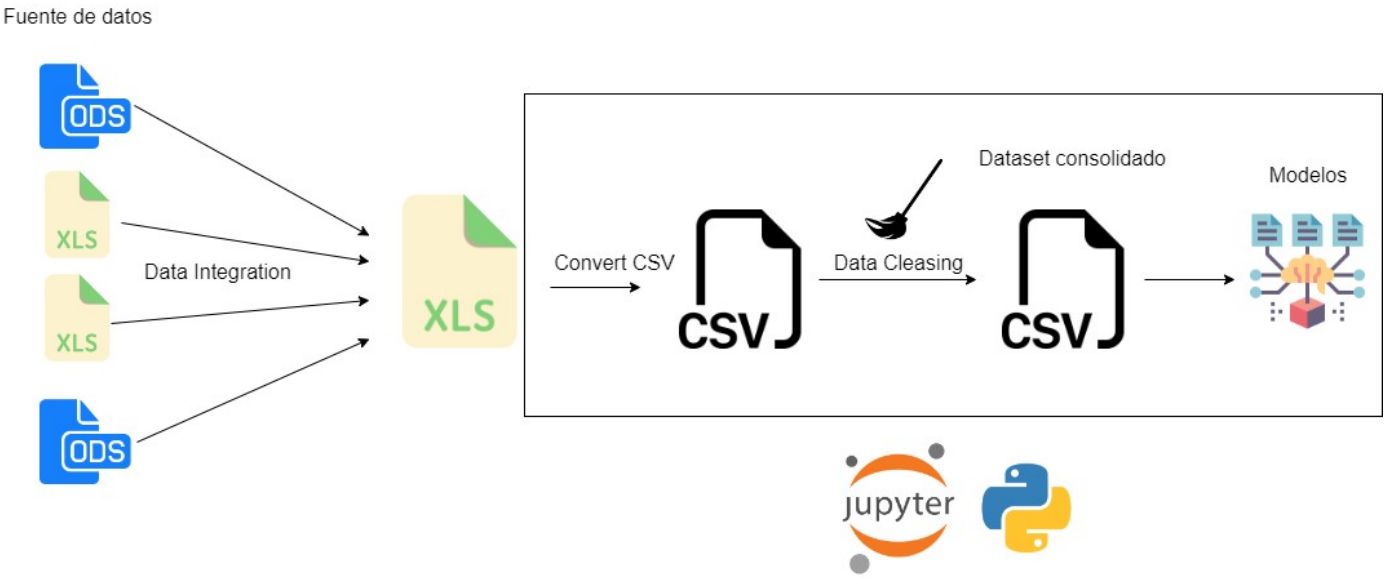


6. Figura cedida por la Junta de Andalucía. Cuenca de Fuente de Piedra.

3.2 Diseño

3.2.1 Diseño del Flujo de Trabajo

Se define en la figura 7 cuál será el flujo de trabajo para llegar a tener un dataset consolidado con el que posteriormente crearemos nuestros modelos.



7. Diagrama flujo de trabajo para la elaboración de un dataset integrado y limpio. Elaboración propia con Diagrams [15].

3.2.2 Detalles del Flujo de Trabajo

Como se puede apreciar en el diagrama 7, partiremos de todos los ficheros para crear un único fichero Excel XLS con todos los datos integrados, es decir, desarrollaremos la fase de integración.

Cuando tengamos este fichero, como se ha comentado anteriormente XLS no es un formato adecuado para trabajar, así que generamos un fichero CSV. El fichero CSV nos servirá como staging área de la parte de limpieza de datos.

La limpieza de datos (*data cleaning*) consiste en encontrar valores nulos (*missing values*), datos duplicados, inconsistencias o incoherencias, incluso encontrar valores atípicos (*outliers*), que deberán analizarse, para validar si se guardan en el dataset.

Para prepocesar los datos se utiliza Jupyter Notebook. Como resultado obtendremos un dataset, que podremos exportar a formato CSV, con una dimensionalidad N; al que aplicaremos, entre otros aspectos, un análisis descriptivo y técnicas PCA.

Nos encontraremos en estas dos fases (integración y limpieza) una casuística muy grande, con muchos casos que resolver, estos casos se detallaran a continuación en la Implementación.

3.3 Implementación

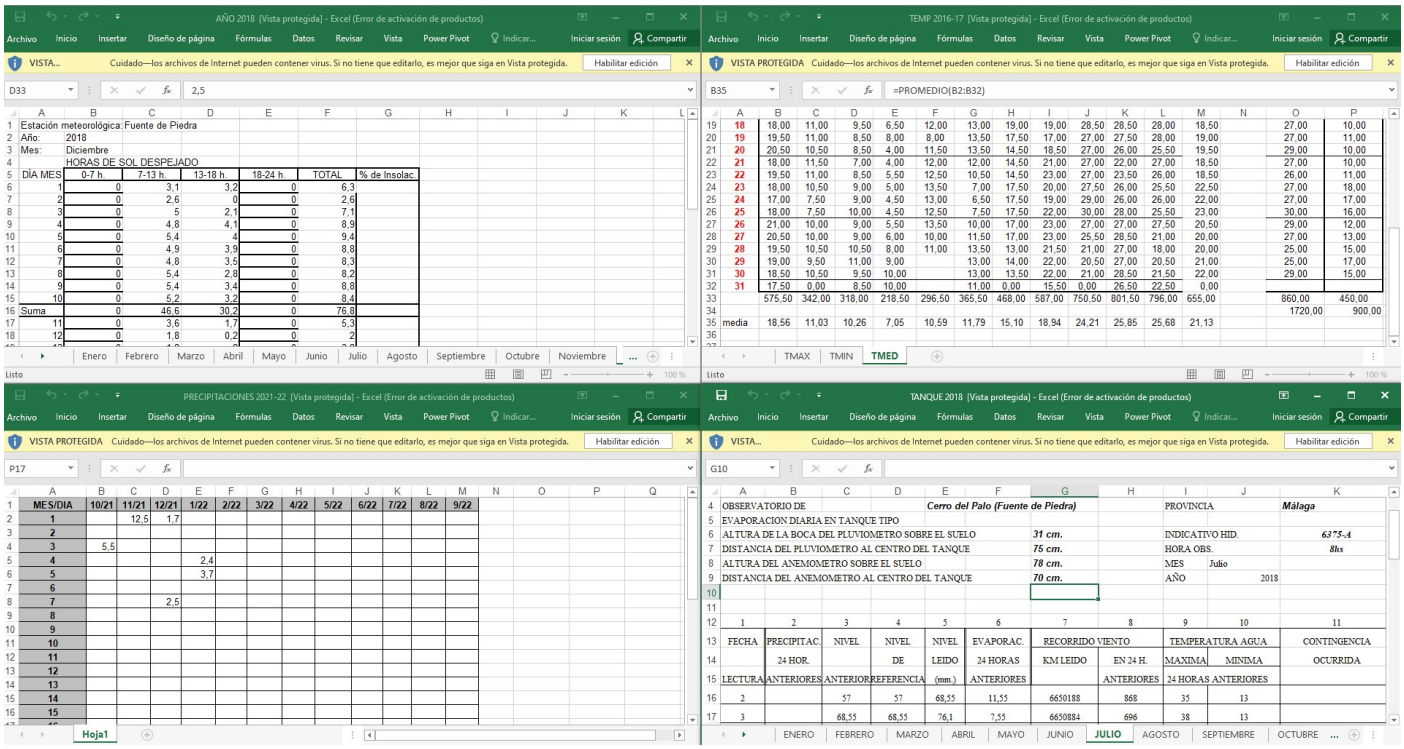
3.3.1 Captura

Históricamente la generación de datos en la Junta de Andalucía ha sido manual y actualmente los sigue siendo. A excepción de alguna estación automática.

La captura de datos se hará de forma manual, los datos son obtenidos de la Junta de Andalucía y, por lo tanto, son ellos los que nos han proporcionado todos los ficheros y anotaciones para su exploración. La variedad de formato en las fuentes se puede apreciar en la imagen 8.

3.3.2 Granularidad

Un aspecto importante para fusionar o integrar los ficheros es la granularidad y cómo relacionar la información. Para nuestro proyecto, la granularidad está a nivel de día/mes/año (dd/MM/YYYY), ya que todos los registros suelen apuntarse manualmente, pocos son generados automáticamente. Es decir, obtendremos para un día en concreto, cuántas horas de Sol hacía, cuáles eran los niveles de los pozos o cuántas aves había en nuestra laguna.



8. Formatos de los ficheros

3.3.3 Integración

Para la integración de los datos se hará de manera manual dada la heterogeneidad de los ficheros y la estructuración individual de cada fichero.

Se ha analizado el conjunto y se han hallado las siguientes características:

Nombre	Años	Formato Granuralidad	Registros	Tipo	Otros
Aves - Censo	[2003 · 2022]	dd/MM/YYYY	87	ODS	
Aves - Reproducción Invernada	[2004 · 2022]	YYYY	52	ODS	
Aves - Parejas y Pollos	[1984 · 2022]	YYYY	38	ODS	
Calidad Agua - Analítica	[2002 · 2011]	dd/MM/YYYY		ODS	- Pocos Datos - Nulos
Calidad Agua - Índices B	[2020]	dd/MM/YYYY		XLS	- Pocos Datos
Calidad Agua - Parámetros B	[2010 · 2020]	dd/MM/YYYY		ODS	- Pocos Datos
Calidad Agua - FQ	[2009 · 2022]	dd/MM/YYYY	10.030	XLS	- No hay Datos para todos los meses del año. - Hay Datos del Arroyo y de la Laguna
Meteorología - Sol	[2005 · 2022]	dd/MM/YYYY	5.960	XLS y ODS	
Meteorología - Lluvia	[1994 · 2022]	dd/MM/YYYY	≈ 9.125	XLS	
Meteorología - Tanque Evaporación	[1995 · 2022]	dd/MM/YYYY	≈ 9.497	XLS	
Meteorología - Temperatura	[1994 · 2022]	dd/MM/YYYY	≈ 9.490	XLS	
Meteorología - Viento	[1995 · 2012]	dd/MM/YYYY	≈ 6.150	XLS	- Muchos datos que ponen “averiado” entre año 2008/2012
Nivel de Laguna	[1983 · 2022]	dd/MM/YYYY	≈ 13.870	XLS	- Datos de anillamiento de aves en una página.
Piezometría	[1983 · 2022]	dd/MM/YYYY	≈ 41.392	XLS y ODS	

Tendremos que definir el dataset destino como un juego de datos completo, es decir que tendremos años completos, el año 2022 no se contemplaría porque solo tenemos informado enero.

Como hay algunos datasets originales que tienen muy pocos datos o incluso muchos valores nulos, hablamos de datos muy pobres en cantidad y calidad, descartaremos integrar a Aves y Calidad del Agua. Dentro de meteorología tenemos al sistema que mide el viento, aunque no podemos utilizarlo porque solo hay datos hasta 2012.

Nos centraremos en los últimos años para construir nuestro dataset, ya que mientra más antiguos son los datos más pobres son.

3.3.4 Preprocesado

En esta etapa, una de las más importantes en un proyecto de Ciencia de Datos, elaboramos el procedimiento por el cuál obtendremos un dataset completo, limpio y depurado.

Para ello utilizaremos Python y Jupyter Notebooks junto con librerías que se podrán consultar en el código.

Como la variedad de formatos es grande y personalizado para cada fichero, vamos a eliminar las cabeceras y estilos manualmente para que el script que vamos a desarrollar pueda procesar más sencillamente los ficheros.

La implementación será en Python con el objetivo de, aprovechando un formato tabular sencillo, donde las columnas son los meses y la filas los días; consigamos construir un fichero nuevo con el histórico. Así conseguiremos convertir la gran cantidad de ficheros en solo unos cuantos que podremos integrar rápidamente.

Nuestro script se aplicará a todos aquellos ficheros que tengan datos suficientes para ser estudiados. Solo algunos ficheros tienen este formato tabular sencillo, el resto se ajustan a mano. Una vez tengamos todos los históricos de cada sección (horas de sol, lluvia, nivel de la laguna, …) construiremos el dataset final con los ficheros que genera el script. Como producto obtenemos lo que se muestra en la figura 9.

Como podemos apreciar, pasamos a la siguiente tarea, a la de la limpieza, de aquellos registros que no sean correctos, de tratar los decimales (puntos por comas por ejemplo), normalización de datos o imputar valores a lluvia con 0 cuando no ha llovido.

	A	B	C	D	E	F	G
1	FECHA	horas_sol	lluvia	max_temp	min_temp	evaporacion	nivel_cm
2	01/01/2017	7,1		14.0	3.0	2,5	31
3	02/01/2017	5,8		13.0	3.0	1	31
4	03/01/2017	0,2		16.0	3.0	0,75	31
5	04/01/2017	4,2		15.0	3.0	2	31
6	05/01/2017	6,2		14.0	3.0	0,05	31
7	06/01/2017	8		15.0	2.0	0,65	31
8	07/01/2017	7,9		14.0	2.0	1,2	31
9	08/01/2017	7,5		15.0	2.0	1,8	31
10	09/01/2017	7,3		14.0	-2.0	2,6	31
11	10/01/2017	2,1		12.0	-1.0	0,05	31
12	11/01/2017	5,8		15.0	1.0	0,75	31
13	12/01/2017	7,5		17.0	3.0	0,9	31
14	13/01/2017	3,3		12.0	3.0	1,3	31
15	14/01/2017	7,7		12.0	0.0	0,7	31
16	15/01/2017	8,2		12.0	-1.0	2,2	31
17	16/01/2017	7,3		13.0	-2.0	1,8	31
18	17/01/2017	8,2		13.0	-2.0	1,3	31
19	18/01/2017	0		14.0	-1.0	1,2	31
20	19/01/2017	0,8		17.0	-1.0	1,4	31
21	20/01/2017	2,8		7.0	1.0	0,85	31
22	21/01/2017	5		9.0	-1.0	0,65	31
23	22/01/2017	7,1		12.0	-1.0	1	31
24	23/01/2017	6,9		12.0	-2.0	1,65	29
25	24/01/2017	7,4		12.0	-3.0	1,5	29
26	25/01/2017	7,5		12.0	-3.0	1,15	29
27	26/01/2017	1	3.0	12.0	-1.0	1,4	29
28	27/01/2017	3,1	6.0	11.0	1.0	1,85	27
29	28/01/2017	6,2		12.0	4.0	0,95	27
30	29/01/2017	6,7		15.0	3.0	1,3	29
31	30/01/2017	7,3		17.0	3.0	1,7	29
32	31/01/2017	3,9		16.0	4.0	1,6	29
33	01/02/2017	3,7		16.0	5.0	2,3	29
34	02/02/2017	0		16.0	6.0	0,7	29
35	03/02/2017	1,1	0,4	15.0	5.0	0,4	29
36	04/02/2017	1,2	1,4	15.0	8.0	2	29
37	05/02/2017	5,1		14.0	9.0	1,2	29
38	06/02/2017	7,2		15.0	4.0	3,05	29

9. Dataset final sin limpieza y tratamiento

3.3.5 Limpieza y tratamiento

En nuestro dataset existen nulos, como se aprecia en la figura 10. Vamos a tratar cada caso, por cada columna tiene que haber 2920 valores:

1. Lluvia: este es el caso más sencillo, dado que la omisión de un valor significa que no ha llovido, por lo tanto el valor será 0.
2. Min_temp: en este caso, solo tenemos un registro nulo, solo hay un registro (10/08/2016), por lo que para simplificarlo haremos la media para hallar el valor de ese registro.
3. Evaporación: los registros que tienen nulos son:
   - 20/03/2013, según la Junta “El nivel supera el -0,5. Quitar agua”

- 14/09/2019, según la Junta por agua.

En este caso, aplicaremos la media, al ser pocos registros.

4. Cm_nivel: el nivel de la laguna medido en centímetros tiene muchos más nulos (296) y es porque no funciona bien el limnígrafo, según nos indica la Junta: “significa que el lignígrafo no escribió en el papel o porque no se ha podido entrar al lignígrafo a darle cuerda”.

   En este caso podemos imputar los valores según KNNImputer que es una biblioteca para la imputación de valores utilizando el resto de datos con el algoritmo de agrupamiento K-Nearest Neightbours.

```
df.info()
```

```
RangeIndex: 3287 entries, 0 to 3286
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   FECHA        3287 non-null   object 
 1   cm_nivel     2991 non-null   float64
 2   horas_sol    3287 non-null   float64
 3   lluvia       1032 non-null   float64
 4   evaporacion  3285 non-null   float64
 5   max_temp     3287 non-null   float64
 6   min_temp     3286 non-null   float64
```

```
df.isnull().sum()
```

```
FECHA            0
cm_nivel       296
horas_sol        0
lluvia        2255
evaporacion      2
max_temp         0
min_temp         1
dtype: int64
```

10. Nulos en el dataset. Elaboración propia.

En nuestro dataset no existen duplicados, ya que hay un registro por cada día del año. Aunque sí existen valores de otro tipo de datos en una variable que es decimal, por ejemplo; “no funciona” en vez de un valor decimal. En este caso, que es mínimo utilizamos la media para la imputación de valores.

El resultado de la limpieza y tratamiento se puede apreciar en la figura 11.

	FECHA	cm_nivel	horas_sol	lluvia	evaporacion	max_temp	min_temp
3282	27/12/2021	33.0	1.0	1.3	-0.25	15.0	12.0
3283	28/12/2021	32.0	2.8	0.5	-0.05	16.0	12.0
3284	29/12/2021	28.0	6.0	0.0	0.50	18.0	5.0
3285	30/12/2021	27.0	7.1	0.0	0.45	23.0	5.0
3286	31/12/2021	26.0	6.6	0.0	0.35	23.0	5.0

11. Dataset procesado.

3.3.6 Descripción del dataset

- **Fecha**: la fecha en la que se capturan los datos, la mayoría son anotaciones de la persona que monitorea la actividad de la laguna. Tipo de dato: cadena de caracteres.

- **Cm_nivel**: el nivel de la laguna se registra de forma automática en un lignígrafo instalado en un pozo en el centro de la misma (centímetros). Tipo de dato: decimal.

- **Horas_sol**: horas de sol despejado en la laguna. Tipo de dato: decimal.

- **Lluvia**: la cantidad de agua que ha llovido, expresada en litros por metro cuadrado. Tipo de dato: decimal.

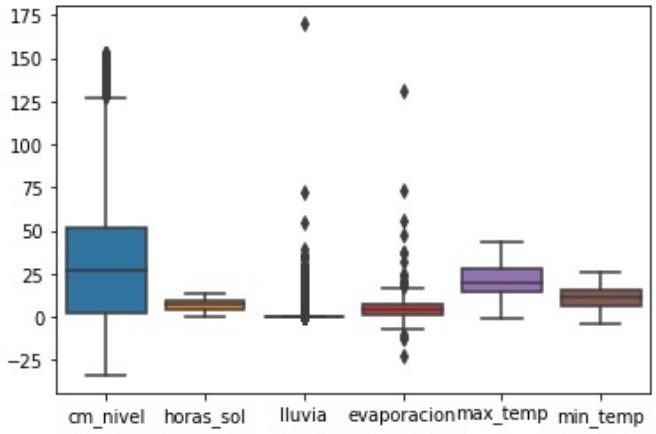
- **Evaporación**: mide la evapotranspiración del agua. Tipo de dato: decimal.

- **Max_temp**: temperatura máxima registrada. Tipo de dato: decimal.

- **Min_temp**: temperatura mínima registrada. Tipo de dato: decimal.

Aunque hay que tener en cuenta que hay variables que presentan valores extremos (outliers), como se ilustra en el diagrama de bigotes 12. El atributo lluvia es normal que tenga una caja plana y también que haya una distribución de outliers que fluctúen porque no llueve proporcionalmente. La evaporación sí que posee más ruido, valores que se salen de los extremos superiores e inferiores que pueden ser por temporadas más calurosas.

Las variables independientes que más correlación tienen con el nivel de la laguna son las horas del Sol y la lluvia. Aunque la mayoría de las variables no tienen una correlación muy alta, como podemos consultar en la matriz de correlación en el NoteBook.



12. Diagrama de bigotes

3.3.7 Métodos de Aprendizaje Automático

A petición de la Junta de Andalucía, se necesita predecir el nivel de la laguna y que las predicciones sean un valor continuo. Es decir que tenemos que aplicar técnicas de regresión para predecir la variable cm_nivel.

Se podría haber optado por discretizarla y predecir una clase categórica, aunque no es lo que se necesita como se ha indicado.

Para poder generar un modelo, necesitamos dividir el conjunto original en dos conjuntos. Conjunto de entrenamiento con el que nuestro modelo aprende a predecir y el conjunto de test, con el que validaremos lo bueno que es prediciendo.
Hay que destacar que el conjunto de datos contempla los años del 2013 al 2021. Se ha probado distintos rangos continuos de años y ampliar el conjunto algunos años no supone una mejora sustancial, sino que incluso puede perjudicar el resultado de los modelos, dado el ruido que añaden.

3.3.7.1 Conjunto de entrenamiento y de test

1. **Conjuntos**

   Para la elaboración de estos conjuntos utilizaremos una función de la librería Scikitlearn, que nos permite hacer la división. Utilizaremos un 20% de datos para el conjunto de validación y un 80% para el entrenamiento. Esta función es train_test_split que devuelve cuatro listas, dos con las variables independientes (X_train y X_test) y dos con las variables dependientes (y_train e y_test).

2. **Normalización**

   Para algunos algoritmos es necesario que la entrada esté normalizada, entiéndase por normalizada que todas las variables del conjunto estén en la misma escala. Así evita crear un sesgo y no asociar a un atributo un peso mayor incorrectamente.

   Para ello, utilizaremos la normalización Min-Max, que utiliza el máximo y mínimo del rango. Los valores están definidos en el rango [0,1]. La figura 13 muestra la fórmula que usaremos.

$$z = \frac{x \text{ - } \min(x)}{\max(x) \text{ - } \min(x)}$$

13. Min Max [16]

3. **Principal Analysis Component**

   Otra opción que tenemos que valorar es reducir la dimensionalidad de nuestro dataset, que a pesar de no ser muy grande, puede dar buenos resultados. Para ello utilizaremos PCA, que "es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información". Es decir, que pasaremos de tener un dataset con cinco atributos a un dataset con dos atributos o dimensiones que explica igualmente el dataset inicial.

   Una vez explicado el procesamiento previo de los datos, resumimos los juegos de datos y aplicamos aprendizaje automático:

   - Conjunto original sin normalizar (X_train, X_test, y_train, y_test)
   - Conjunto aplicado PCA con dos dimensiones y normalizado (X_train_pca, X_test_pca)
   - Conjunto original normalizado (X_train_normalized, X_test_normalized)

3.3.7.2 Regresión Lineal Múltiple

Definición: La regresión lineal múltiple es una extensión de la regresión lineal simple [17], ya que se necesita más de una variable predictora para predecir la variable objetivo [18].

Aplicación: entrenamos tres modelos de regresión lineal para la variable objetivo que es nivel de la laguna (cm_nivel) y medimos su precisión con el score de R cuadrado.

Precisión:

- Coeficiente de determinación R^2: **0.46**
- Coeficiente de determinación R^2 con PCA: **0.22**
- Coeficiente de determinación R^2 normalizada: **0.46**

3.3.7.3 Random Forest Regressor

Definición: Random Forest Regressor [19] es un algoritmo basado en árboles de decisión que constituyen un bosque y que unifica los resultados de cada árbol para obtener un resultado mejor [20].

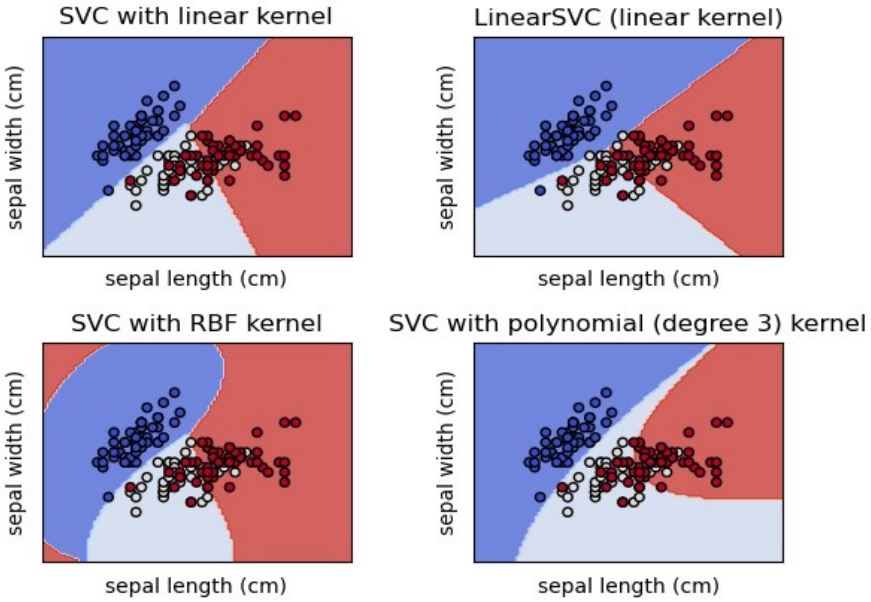
Aplicación: primeramente, utilizamos validación cruzada, que explicaremos en el punto 3.3.7.8 para hallar los mejores hiperparámetros. Además, entrenamos tres modelos de Random Forest para la variable objetivo que es nivel de la laguna (cm_nivel) y medimos su precisión con el score de R cuadrado.

Precisión:

- Coeficiente de determinación R^2: **0.607**
- Coeficiente de determinación R^2 con PCA: **0.236**
- Coeficiente de determinación R^2 normalizada: **0.604**

3.3.7.4 Regresión de Soporte Vectorial

Definición: SVR tiene como objetivo seleccionar un hiperplano óptimo de separación, es decir, que tenga un número máximo de puntos. El concepto ilustrativo está en la figura 14. [21] [22]



14. SVC que es similar a SVR pero aplicado a la clasificación [23].

Aplicación: primeramente, utilizamos validación cruzada para hallar los mejores hiperparámetros. Además, entrenamos tres modelos de SVR para la variable objetivo que es nivel de la laguna (cm_nivel) y medimos su precisión con el score de R cuadrado.

Precisión:

- Coeficiente de determinación R^2: **0.577**
- Coeficiente de determinación R^2 con PCA: **0.219**
- Coeficiente de determinación R^2 normalizada: **0.571**

3.3.7.5 Regresión KNN

Definición: este algoritmo es el K-vecinos más cercanos aplicado a la regresión. El objetivo es predicho por la interpolación [24].

Aplicación: primeramente, utilizamos validación cruzada para hacer tuning. Además, entrenamos tres modelos de KNN Regresión para la variable objetivo que es nivel de la laguna (cm_nivel) y medimos su precisión con el score de R cuadrado.

Precisión:

- Coeficiente de determinación R^2: **0.582**
- Coeficiente de determinación R^2 con PCA: **0.251**
- Coeficiente de determinación R^2 normalizada: **0.553**

3.3.7.6 Multivariate Adaptative Regression

Definición: MARS se define como “método de regresión flexible que busca automáticamente interacciones y relaciones no lineales. Los modelos terrestres se pueden considerar como modelos lineales en un espacio base dimensional superior. Cada término en un modelo de la Tierra es un producto de las llamadas "funciones de bisagra". Una función bisagra es una función que es igual a su argumento donde ese argumento es mayor que cero y es cero en todo lo demás.” [25] [26]

Aplicación: En este caso utilizaremos una librería externa que es pyearth.earth. También utilizamos validación cruzada para hallar los mejores hiperparámetros. Además, entrenamos tres modelos de Earth para la variable objetivo que es nivel de la laguna (cm_nivel) y medimos su precisión con el score de R cuadrado.

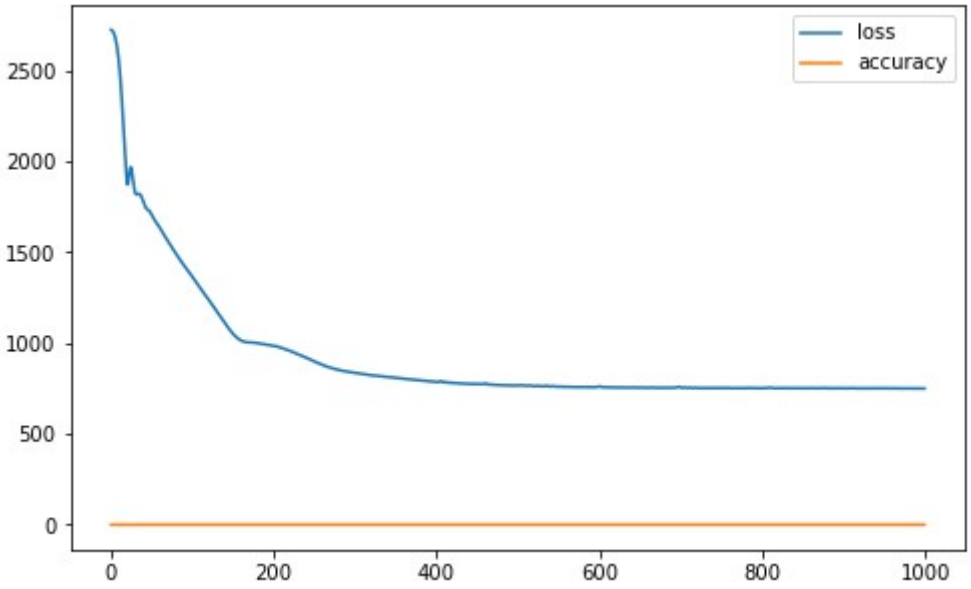
Precisión:

- Coeficiente de determinación R^2: **0.495**
- Coeficiente de determinación R^2 con PCA: **0.291**
- Coeficiente de determinación R^2 normalizada: **0.535**

3.3.7.7 Redes Neuronales Artificiales ANN

Definición: Una red neuronal es una estructura compleja definida por una capa de entrada, N capas ocultas y una capa de salida, que suele ser la variable a predecir. Cada capa está compuesta por neuronas que están conectadas y procesan datos aplicando fórmulas. Son capaces de aplicar técnicas de clasificación y regresión, aunque consumen muchos más datos que el resto de algoritmos de aprendizaje automático para que sean precisas [27].

Aplicación: Para crear la red, utilizaremos una primera capa con 4 neuronas, dos ocultas y una capa de salida con una neurona. Compilamos el modelo y apreciamos que el loss decae, es decir el error entre la salida y el valor real es cada vez menor, aunque el valor de la precisión (accuracy) es constante. La precisión de la red es muy baja, por debajo del 5 por ciento. Consultar figura 15.



15. Gráfica de pérdida y precisión

Precisión: < 5%

3.3.7.8 Validación Cruzada (CrossValidation)

Aplicaremos validación cruzada con la función GridSearchCV de la librería Scikit Learn. [28]

Esta función calculará, dado un grid (configuración de parámetros), cuál es la combinación más óptima de hiperparámetros para nuestros modelos. Cada modelo de aprendizaje automático tiene unos hiperparámetros que se pueden apreciar en los grid definidos en el código Python. Para ello hemos creado una función en Python que llama a esta y además calcula el tiempo de ejecución.

4. Conclusiones

4.1 Conclusiones

En la Laguna de Fuente de Piedra, según los gráficos de dispersión elaborados con la librería Seaborn, apreciamos que la tendencia de la evaporación de agua aumenta cuando aumenta la temperatura y también cuando aumenta las horas de sol sobre el entorno.

Además cuando llueve, el ambiente es más húmedo y la evaporación es menor. El nivel de la laguna aumenta ligera y suavemente. Esto es así, porque la extensión de la laguna es grande.

En cuanto a los modelos de regresión, en general todos tienen una bondad parecida, es decir predicen con una precisión similar, excepto las redes neuronales. Como es lógico, el conjunto de datos es pequeño para entrenar una red neuronal artificial, ya que requieren de grandes volúmenes de datos, por eso la precisión es baja.

En general, a pesar de que la normalización es un punto a favor para los modelos de aprendizaje automático, los resultados son similares escalando los datos. Además, podemos destacar que en este caso la técnica de PCA con dos dimensiones no aporta propiedades beneficiosas a los modelos, concluyen mejores resultados las versiones anteriores.

4.2 Discusión

El proyecto ha sido un desafío, sobre todo en la parte de integración de datos. Las fuentes de datos estaban muy dispersas, cada dataset tiene una estructura única, hay que seleccionar cuáles son los datos relevantes, la mayoría no siguen una estructura tabular limpia, las columnas no siguen un tipo de datos específico, etc.

Otro aspecto importante ha sido la carencia de datos, no hemos tenido un conjunto amplio, esto hace que los modelos no aprendan tanto como lo harían con más datos. Podemos afirmar que hemos conseguido el OP, objetivo principal. Aunque no los secundarios S1,S2 y S3 debido a la incompletitud y pocos datos que tenemos.

La aplicación de varios algoritmos potencia hallar un método óptimo para elaborar nuestro modelo. También se ha aplicado una herramienta muy interesante en el campo de Ciencia de Datos; la búsqueda de hiperparámetros que hace que nuestro modelo RandomForestRegressor consiga un 60% de precisión. El bosque de árboles no consigue un resultado excelente, pero tampoco bajo.

Estos resultados se podrían mejorar automatizando la generación de los datos en el origen. Ya sea con sensores u otros mecanismos, que estos sean los que generen los datos y se almacenen en una base de datos y no en ficheros que se generan manualmente. Además, se plantea, en caso de que no se pueda llevar a cabo lo comentado anteriormente, establecer una plantilla para que se siga un estándar tanto en formato como en tipos de datos.

De esta manera evitamos introducir errores y aumentar la calidad del dato, además de ahorrar en tener que estar midiendo constantemente manualmente todos los parámetros.

5. Trabajo Futuro

A continuación, se proponen los trabajos a futuro para la mejora o ampliación del proyecto.

Comenzamos por la ampliación del conjunto de datos, sería interesante integrar distintas fuentes ya sean internas de la Junta de Andalucía o externas, así enriquecemos, hacemos más sofisticados los modelos y podremos hallar nuevos descubrimientos.

Tenemos que tener en cuenta que podemos ampliar el número de hiperparámetros que GridSearchCV utilizará para crear las combinaciones. Si tenemos presupuesto, podríamos aumentar las prestaciones de nuestra máquina tradicional o con servicios en la nube y conseguir mejores resultados.

Se recomienda como algo relevante, estudiar/automatizar las fuentes de datos, para tener un dato de más calidad que no induce a errores. Completadas las carencias anteriores de formato, tipos de datos y automatización, cumplir con los objetivos secundarios. Por ejemplo, para predecir el número de ejemplares de Flamenco.

Tras un intento por correo, sin éxito, de saber cómo calcular el nivel freático de la cuenca y estudiar si tenemos los suficientes datos con IGME(Instituto Geológico y Minero de España), sería interesante plantear reuniones para elaborar un conjunto de datos ideal para predecirlo como trabajo futuro. Así hallamos en función, de los elementos de la cuenca (ríos, pozos, etc) si habrá suficiente agua subterránea y si se puede seguir extrayendo agua.

6. Glosario

Limnígrafo: Un limnígrafo es un mecanismo que mide el nivel del agua, está formado por un flotador unido a una plumilla que marca el nivel en un papel fijado a un tambor, que gira mediante un mecanismo de relojería, generando una gráfica de niveles contra el tiempo conocida como limnograma [11].

Piezometría: Parte de la hidrología que estudia los métodos para determinar la cantidad de agua (subterránea) existente en un lugar sobre una capa impermeable de un terreno. Se entiende por cota o nivel piezométrico, a la altitud o profundidad (en relación a la superficie del suelo), del límite entre la capa freática y la zona vadosa en un acuífero. Este nivel se mide usando un piezómetro [12].

Heliógrafo: Instrumento que registra los periodos de radiación solar suficientemente intensa para producir sombras definidas [13].

Tanque de evaporación: Es el instrumento que se emplea para medir la evaporación del agua en la atmósfera [14].

7. Limitaciones

Se detallan las limitaciones o restricciones que se ha tenido durante el desarrollo del trabajo.

Calidad del dato:

1. Unicidad: en nuestro caso hemos aprovechado aquellas fuentes que nos permiten construir un dataset con una granularidad definida por fecha. Esto hace que cada observación sea única. Se destaca que algunos ficheros que no se han contemplado para la creación del dataset definitivo contenían nulos.

2. Completitud: el juego de datos que contempla a todos los datasets iniciales no está completo. Es decir, hay ficheros que no contienen datos para todos los años o para todas las observaciones.

3. Uniformidad: la estructura de los ficheros difiere bastante entre ellos, además mientras más antiguos son, menos uniformidad.

Volumetría de datos:

Nuestra fuente tiene muchos ficheros, aunque no tienen un volumen grande de datos.

8. Código

El código desarrollado en Notebook de Jupyter con Python puede encontrarse en la entrega de la PEC3 o en este repositorio;

https://github.com/miguel-a-ngel/Machine-Learning-applied-to-enviroment-Fuente-de-Piedra-Lake

9. Bibliografía

1. ¿Qué vas a diseñar?. (2022).*Canva*. https://www.canva.com/design/play?category=tADWs1ZQ3GI
2. Ventana Del Visitante. (2022). *Laguna de Fuente de Piedra*. https://www.juntadeandalucia.es/medioambiente/portal/web/ventanadelvisitante/detalle-buscador-mapa/-/asset_publisher/Jlbxh2qB3NwR/content/laguna-de-fuente-de-piedra-8/255035
3. Ziyad Sami, B.F., Latif, S.D., Ahmed, A.N. et al. (2022). *Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan*. https://doi.org/10.1038/s41598-022-06969-z
4. Copernicus. (2022). *EU-Hydro - River Network Database*. https://land.copernicus.eu/imagery-in-situ/eu-hydro/eu-hydro-river-network-database?tab=download&selected:list=eu-hydro-guadalquivir-fgdb
5. Mohamad Javad Alizadeh Mohamad Reza Kavianpour. (2015). *Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean*. https://doi.org/10.1016/j.marpolbul.2015.06.052
6. Towards Data Science. (2020). *Using machine learning to predict Rhine water levels*. https://towardsdatascience.com/using-machine-learning-to-predict-rhine-water-levels-44afce697074
7. Han, X., Guo, Y., Mi, C. et al. (2017). *Machine Learning Model Analysis of Breeding Habitats for the Black-necked Crane in Central Asian Uplands under Anthropogenic Pressures*. https://doi.org/10.1038/s41598-017-06167-2
8. Benjamin Kellenberger, Thor Veen, Eelke Folmer, Devis Tuia. (2021). *21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning*. https://doi.org/10.1002/rse2.200
9. S. E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic y S. Alessandrini. (2018). *"Machine Learning for Applied Weather Prediction", 2018 IEEE 14th International Conference on e-Science (e-Science)*. https://doi.org/10.1109/eScience.2018.00047
10. Kabir Rasoulia, William W.Hsieha, Alex J.Cannonb. (2012). *Daily streamflow forecasting by machine learning methods with weather and climate inputs*. https://doi.org/10.1016/j.jhydrol.2011.10.039
11. Espinosa, F. S., Olguín, G. Á., & Leyva, F. H. R. (2015). *Diseño y construcción de un limnígrafo electrónico*. https://redib.org/Record/oai_articulo1623164-dise%C3%B1o-y-construcci%C3%B3n-de-un-limn%C3%ADgrafo-electr%C3%B3nico
12. Confederación Hidrográfica del Guadiana. (2022). *Piezometría*. https://www.chguadiana.es/cuenca-hidrografica/hidrologia/aguas-subterraneas/piezometria
13. MeteoGlosario Visual AEMET Diccionario ilustrado de meteorologia. (2022). *Heliógrafo*. https://meteoglosario.aemet.es/es/termino/1021_heliografo

14. Metogalicia. (2022). *Glosario de meteorologia Evaporación*. https://www.meteogalicia.gal/web/informacion/glosario/est25.action?request_locale=es
15. Diagrams.net. (2022). *Diagrams*. https://app.diagrams.net/?src=about
16. Study Machine Learning. (2019). *Machine Learning Tutorials*. https://studymachinelearning.com/feature-preprocessing-for-numerical-features/
17. Medium. (2021). *Multiple Linear Regression Implementation in Python*. https://medium.com/machine-learning-with-python/multiple-linear-regression-implementation-in-python-2de9b303fc0c
18. ScikitLearn. (2022). *sklearn.linear_model.LinearRegression*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression
19. DataScientest. (2022). *Random Forest: Bosque aleatorio. Definición y funcionamiento*. https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento
20. Towards Data Science. (2022). *Random Forest Regression*. https://towardsdatascience.com/random-forest-regression-5f605132d19d
21. Towards Data Science. (2022). *Unlocking the True Power of Support Vector Regression*. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0
22. Medium. (2019). *Support Vector Regression in 6 Steps with Python*. https://medium.com/pursuitnotes/support-vector-regression-in-6-steps-with-python-c4569acd062d
23. ScikitLearn. (2022). *Support Vector Machines*. https://scikit-learn.org/stable/modules/svm.html
24. ScikitLearn. (2022). *sklearn.neighbors.KNeighborsRegressor*. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
25. Contrib Scikit Learn. (2022). *Source code for pyearth.earth*. https://contrib.scikit-learn.org/py-earth/_modules/pyearth/earth.html
26. Towards Data Science. (2020). *MARS: Multivariate Adaptive Regression Splines — How to Improve on Linear Regression?*. https://towardsdatascience.com/mars-multivariate-adaptive-regression-splines-how-to-improve-on-linear-regression-e1e7a63c5eae
27. ScikitLearn. (2022). Neural network models (supervised). https://scikit-learn.org/stable/modules/neural_networks_supervised.html
28. ScikitLearn. (2022). *sklearn.model_selection.GridSearchCV*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html