2023

Electricity Forecast & Analysis Tool (EFAT)

K-SCHOOL - DATA SCIENCE MASTER, 2023 – TFM

JAVIER LAFUENTE ESCALONILLA

Index

1. Introduction

Electricity Forecast and Analysis Tool (EFAT, from now on) was conceived as part of the K-School Data Science (5th streaming edition) final project. The aim of the project is to put all possible technologies and knowledge acquired during the master at the service of achieving a better understanding of electricity market.

During the last years, countless news have been published related to the volatility and uncertainty of the electricity market and its effects on multiple socioeconomic aspects. Its impact transcends the market itself, affecting practically all markets in world. In the end, the ultimate impact of all these consequences reach out to the citizens and companies around the world in multiple ways.

A comprehensive understanding of how this market operates, and perhaps the ability to predict production shifts and price changes before they occur, can offer a significant competitive advantage.

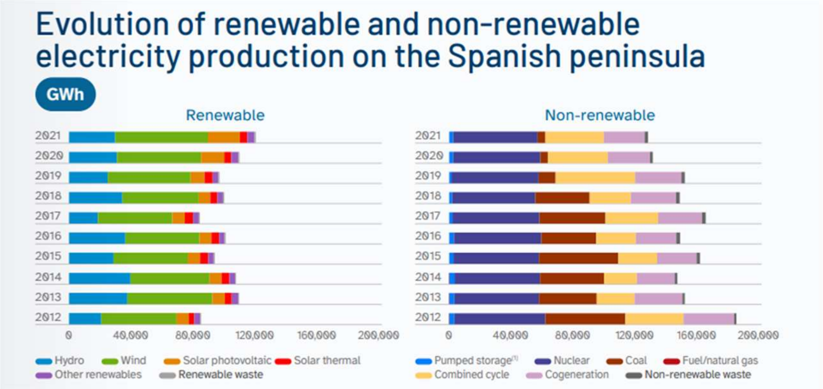
Giving the context –a master's final project- the scope of the project could not be as ambitious as initially dreamed of. There are many variables that can impact on the electricity price or resources availability, and it would be an immense task to resume all these variables into a single project.

However, it is possible to analyze portions of that market in order to extract conclusions that can drive to potential efficiencies or improvements. This objective is the essence of the EFAT Project.

Focusing on renewable energy resources in Spain, EFAT project target is to establish a model that can predict energy generation & demand by analyzing several features related to the weather, day of the week etc. And all of this analyzed per region – Comunidades Autonomas -.

Renewables resources are the future. Understand how it work and being able to predict its behavior may lead to decision-driven conclusions for companies and families:

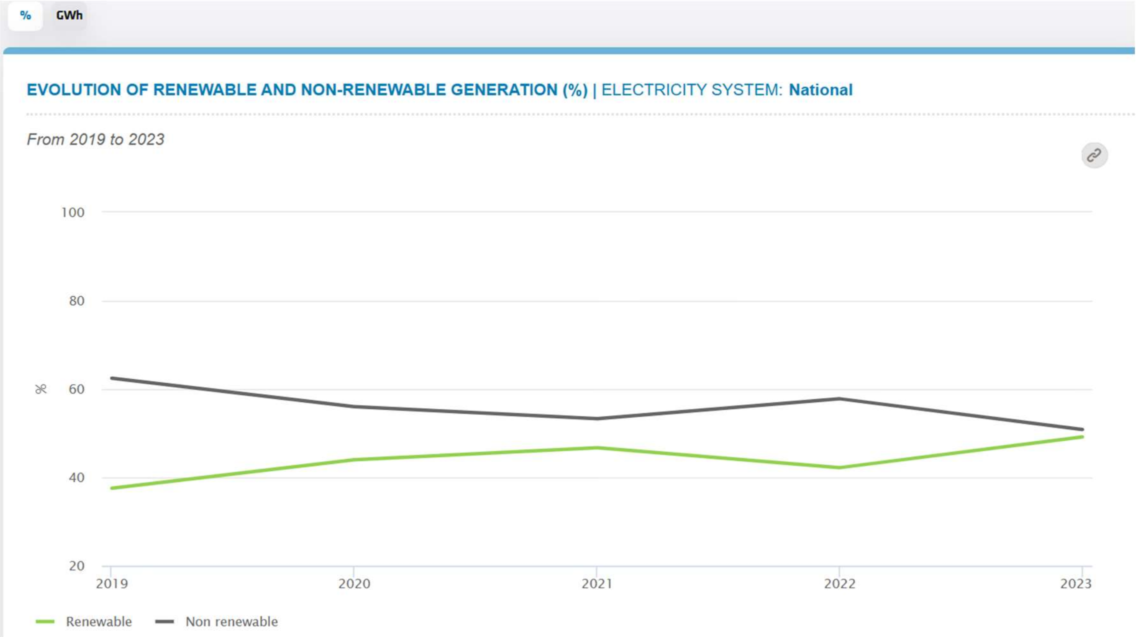
Figure 1:Evolution of electricity technologies in Spain



Note: The Spanish Electricity System Report 2021, p.31, Red Eléctrica de España, 2021

As it can be seen in *Figure 1*, the evolution of the renewable electric resources have been increasing during the last years -from 39% to almost 50% of total production. It is something that can be easily observed by visiting Red Eléctrica de España, where it can be found an interesting dashboard with many information relevant to the project (*Figure 2*):

Figure 2: Evolution of renewable and non-renewable generation in % (2019-2023)



Note: Evolution of renewable and non-renewable generation, ree.es/en/datos/generacion, Red Eléctrica de España

From Figure 1 it can be observed that the main renewables energetic technologies are the following ones: Eolic (wind), Hydro (Hidraulic) and Solar (photo & thermal). Those electricity resources will be the targets to predict within the project.

The Project can be split mainly in five key stages that will be explained further in the document:

- Data Extraction: Data from different public institutions has been processed and analyzed, mainly related to weather forecast and historical data, and electricity generation and demand.
- Data Processing: Several transformations must be applied to the data in order to cross all the information
- Exploratory Data Analysis (EDA): After downloading and processing the data, it's crucial to analyze it for a comprehensive understanding and to pinpoint the most significant variables for the study. The EFAT project was developed in Python, with the EDA mainly conducted using the 'pandas' library, and visualization primarily facilitated mainly by 'Matplotlib'.
- Model Elaboration: Using scikit-learn algorithms, different models were built for each technology (solar, hydraulic, eolic etc.) in order to have the model that fits the best for each one.
- Conclussions & FrontEnd: Frontend build on streamlit is the final purpose of the project, something easily to work with for the general public and which shows the estimations provided by the models for the following days.

During all these processes, several iterations have been performed in order to improve the model or enriching the data.

2. Data Extraction

One of the targets of the design of EFAT project was to make it lightweight. Therefore, there is no need to download and store large files locally. Information is downloaded through a shared folder on Google Drive that is mentioned in the ReadMe document and in the Jupyter Notebooks. All the data and the models are stored there and the EFAT project code just read it online.

Data is the main asset of EFAT Project. It has been -and it is, as it consumes real-time data- extracted from three main resources:

- Agencia Estatal de Meteorología (AEMET) (https://www.aemet.es/es/portada): It brings tones of data from 2014 onwards and it provides also information about the forecasting for the following days regarding the main meteorological variables, such as temperature, sky status or wind speed expected. It also provides an extensive API which helps to extract data from the web in a faster way.
- Red Eléctrica de España (REE, https://www.ree.es/): Also through an API easy to use, it provides information related the electricity generation and demand for different ranges of time, and it allows to search by 'Comunidad Autónoma' or 'Electric System'. It is the key to access to all the information related to electricity.

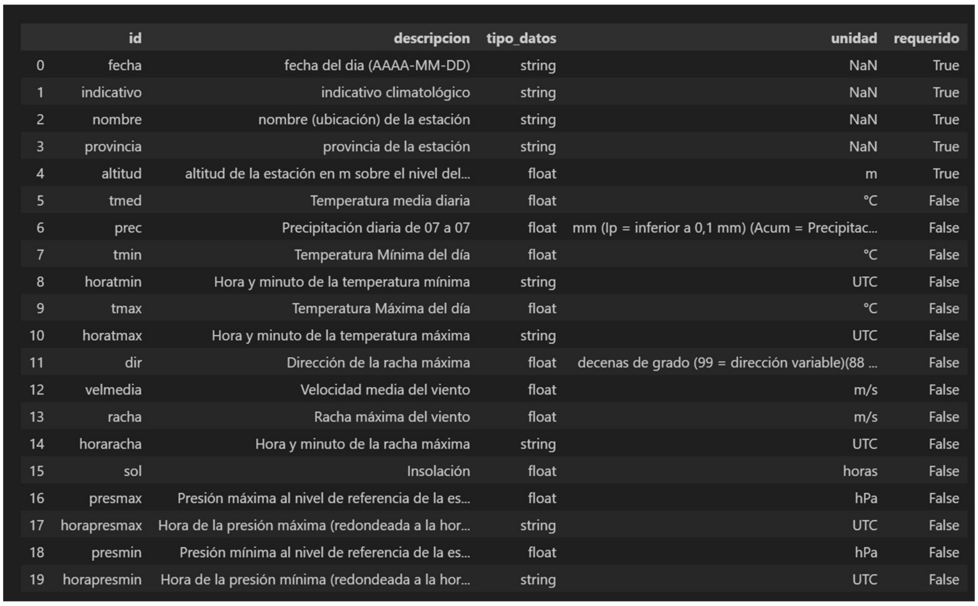
- Ministerio para la Transición Ecológica y Reto Demográfico (MITECO, Ministerio para la Transición Ecológica y el Reto Demográfico (miteco.gob.es)): MITECO is responsable for maintaining and financing studies and analysis related to ecology and spanish environment. In this sense, it has been the key for EFAT project in order to access to the information related to the water reservoir in Spain.

In the different notebooks that composed the EFAT project, the main information that is downloaded from these resources can be resume as follows:

1. From AEMET:

- Data is downloaded through the API provided in: [AEMET OpenData - Agencia Estatal de Meteorología - AEMET. Gobierno de España]() It is necessary to obtain an API key, but it is easy.
- Weather Historical Data: AEMET offers a large amount of historical data coming from the different meteo stations they have around the country. The variables included, which are the variables that are using for training the model and predicting the target, are the ones in Figure 3:

Figure 3: Metadata from AEMET, Historical-Data

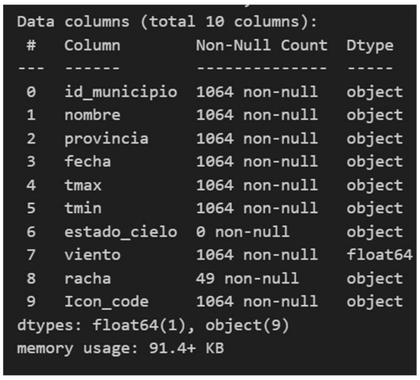


	id	descripcion	tipo_datos	unidad	requerido
0	fecha	fecha del dia (AAAA-MM-DD)	string	NaN	True
1	indicativo	indicativo climatológico	string	NaN	True
2	nombre	nombre (ubicación) de la estación	string	NaN	True
3	provincia	provincia de la estación	string	NaN	True
4	altitud	altitud de la estación en m sobre el nivel del...	float	m	True
5	tmed	Temperatura media diaria	float	°C	False
6	prec	Precipitación diaria de 07 a 07	float	mm (Ip = inferior a 0,1 mm) (Acum = Precipitac...	False
7	tmin	Temperatura Mínima del día	float	°C	False
8	horatmin	Hora y minuto de la temperatura mínima	string	UTC	False
9	tmax	Temperatura Máxima del día	float	°C	False
10	horatmax	Hora y minuto de la temperatura máxima	string	UTC	False
11	dir	Dirección de la racha máxima	float	decenas de grado (99 = dirección variable)(88 ...	False
12	velmedia	Velocidad media del viento	float	m/s	False
13	racha	Racha máxima del viento	float	m/s	False
14	horaracha	Hora y minuto de la racha máxima	string	UTC	False
15	sol	Insolación	float	horas	False
16	presmax	Presión máxima al nivel de referencia de la es...	float	hPa	False
17	horapresmax	Hora de la presión máxima (redondeada a la hor...	string	UTC	False
18	presmin	Presión mínima al nivel de referencia de la es...	float	hPa	False
19	horapresmin	Hora de la presión mínima (redondeada a la hor...	string	UTC	False

- Weather forecast: Include the variables that are used for predicting the renewable electricity generation and demand in the project. The way the API provides the data is different from the historical data, and to achieve to download the data and processing it so it can be used has been one of the main milestones of EFAT project. It only allows to download the prediction for a single ‘municipio’ or municipality each time, so in the notebooks the downloaded information is accessed one by one and stored in a file to be used afterwards.

After downloading the information would look like the dataframe .info() in *Figure 4,* in which Icon_code would be proportional to ‘sol’ in Historical Data:

Figure 4: Weather Forecast dataframe columns after processing



```
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   id_municipio  1064 non-null   object
 1   nombre        1064 non-null   object
 2   provincia     1064 non-null   object
 3   fecha         1064 non-null   object
 4   tmax          1064 non-null   object
 5   tmin          1064 non-null   object
 6   estado_cielo  0 non-null      object
 7   viento        1064 non-null   float64
 8   racha         49 non-null     object
 9   Icon_code     1064 non-null   object
dtypes: float64(1), object(9)
memory usage: 91.4+ KB
```

- ‘Icon_code’ columns, as commented, is related to the expected isolation in the following days. The information was not provided in the same scale than ‘sol’ in Historical-Data, so it is processed to align both data through another file that was directly provided by AEMET by email after contacting them and which relates both files. It can be found on Google Drive shared folder with the name ‘estado_cielo.csv’
- ‘Municipios’ masterdata: contains all municipalities from Spain, and it is used to match it with the provinces -and, in the end, to access to the CCAA- when downloading the weather forecast.
-

2. From REE

- Generation data from 2014 to 2022: Downloaded as a dataframe, it includes the columns:
    - Value: Represents generation in Mw/h[1], which will be the standard for the whole project. In a daily basis, for the entire country.
    - Percentage: percent of the total of the electricity produced each day that each technology represents.
    - Datetime: Date

[1]“A Megawatt Hour is a unit of measurement that describes the amount of energy produced by one Megawatt over the course of one hour”, [Megawatt Hour (MWh) | Meaning, Uses, & Benefits (carboncollective.co)](), Zach Stein

  - Type: technology or energy resource.

- Generation by CCAA: It is useful to have the generation of each technology per 'Comunidad Autónoma', but it is provided in a monthly basis. Thus, it is converted to daily data within the notebooks. This dataset contains the same columns as the daily file but with an additional 'community_name' column.
- PowerInstalled by CCAA: It shows the generation capacity per CCAA for each technology. It comes also in a monthly basis, so it is transformed in the notebooks to daily using 'interpolate'.
- Demand Daily: As with generation information, is provided in a daily basis as a total for the whole country.
- Demand by CCAA: It is transformed into daily and merged with 'demand daily' applying the monthly demand percent for each CCAA to all the days of that month.

All these files are, in the end, the target of the models that are prepared in the project.

3. From MITECO

Estimation of hydraulic production can be affected by many factors, including political ones or others driven by private companies' interests. However, the main factor that can determine the level of hydraulic production in a particular moment in time is the status of the national water reservoir.

MITECO offers an in-depth analysis of the historical and current evaluation of water resources. This includes advanced apps with real-time data. One of the most interesting files available is a database with the historical status of the 'embalses' (dams) in Spain from 1988 to 2022 on a weekly basis (El Boletín Hidrológico semanal (miteco.gob.es))

The file only contains five columns: 'AMBITO_NOMBRE' (watershed to which the 'embalse' belongs), 'EMBALSE NOMBRE', 'FECHA' ,'AGUA_TOTAL' (maximum capacity of the dam), 'AGUA_ACTUAL' (current level) & 'ELECTRICO_FLAG' (a flag that indicates if the dam is used for producing electricity or not). All these columns are useful, but is still needed to relate the dam with the province or the CCAA to connect it with other files.

To do so, MITECO also provides a masterdata file which includes for each 'embalse' information including the province related: Inventario de Presas y Embalses (miteco.gob.es)

4. Others

Apart from these three main sites, other sources have been useful for different purposes along the project:

- Instituto Nacional de Estadística (INE): From INE it was downloaded the mapping table with the 'Provincias' and 'Comunidades Autónomas' codes that are used in AEMET when extracting the data from the API.
- Geopandas map from GADM: In order to plot several variables on a map of Spain with the CCAA divisions included, a shapefile with the information related to latitude and longitude was downloaded.

3. Methodology

The initial phase of the process, detailed below, was carried out using smaller dataframes for each technology, focusing solely on data from 2014. All related notebooks from this phase can be found in the project's shared Google Drive folder. Subsequently, insights drawn from this preliminary phase were employed to develop the 'EFAT Model & Conclusions' Notebook, which now serves as the primary code source for the project preprocessing and visualization, and 'EFAT_Predictions' which is the notebook that tests the models with the latest weather information available.

1. Preprocessing

It is, by far, the largest and most time-consuming phase of the EFAT Project. It is included here also the phase consisting in acquiring all the files needed to process the information and get to some valuable points. But once that part is done, it has been a continuous process of cleaning, adjusting and modifying files to finally have a complete dataframe with all the information needed.

Using pandas library, the idea was to bring all the information together by merging generation and demand files from REE, weather historical data from AEMET and 'embalses' information from MITECO. During the process, some of the transformations that are done for several files are:

- Feature engineering: This involved converting dataframes from a weekly to daily frequency, rectifying datetime column formats, pivoting categorical columns, and extracting key features such as the percentage representation for CCAA generation.
- Handling NA values: Weather data posed challenges with several NA values. Most of the times the impact was reduced to almost none when grouping by

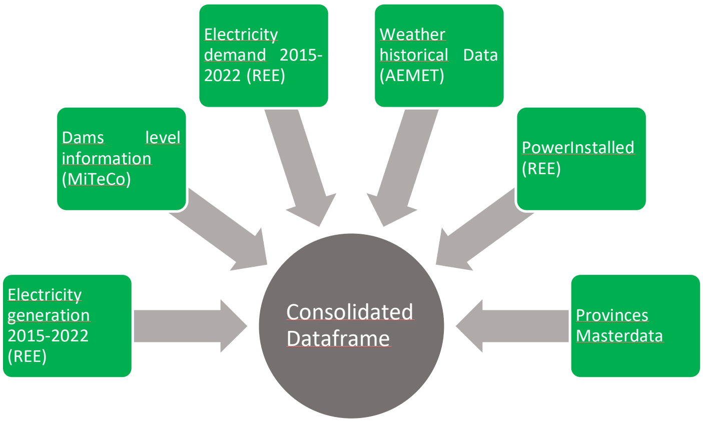
CCAA or date. For other cases, In general 0 values have been consider, unless in case of 'sol' values for example in which the number of 0 values were not meaningful and the rows were dropped.

- Merging files: A key challenge was ensuring data compatibility across files. This often required manipulating multiple files simultaneously to ensure consistency in fields such as 'comunidad_autonoma', 'fecha', and periodicity of the data.

From something bigger to something smaller: In the end a dataframe called 'consolidated_df' is achieved by merging files, but during the process several features that were not necessary have been removed and the information is shown at CCAA level, easier to analyze.

In the end, the sum up of the files used for the final consolidated file would be as follows:

Figure 5: How dataframe consolidated is created, EFAT_project


Electricity demand 2015-2022 (REE)
Weather historical Data (AEMET)
Dams level information (MiTeCo)
PowerInstalled (REE)
Electricity generation 2015-2022 (REE)
Consolidated Dataframe
Provinces Masterdata

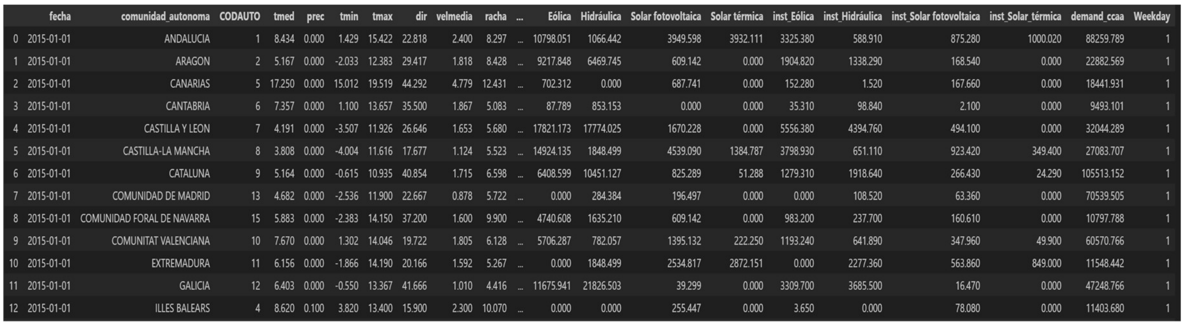
2. Exploratory Data Analysis

While EDA could be viewed as an extension of the preprocessing phase, it merits its distinct section, given its critical role in understanding and refining data.

This point englobes all the analysis of the variables included in the files incorporated in the model. With Pandas methods (head, info, describe etc.) it Is possible to have a quick first view of the data and how its structured. It has been necessary also to use unique() methods to see how categorical values were written, analyze or change several dtypes, removing unused points or commas on strings, decoding text for several columns as each one came with different criteria etc.

In the end, the dataframe cleaned would be the one on Figure 5.

Figure 6: Consolidated Dataframe including all the variables for the study, EFAT Project



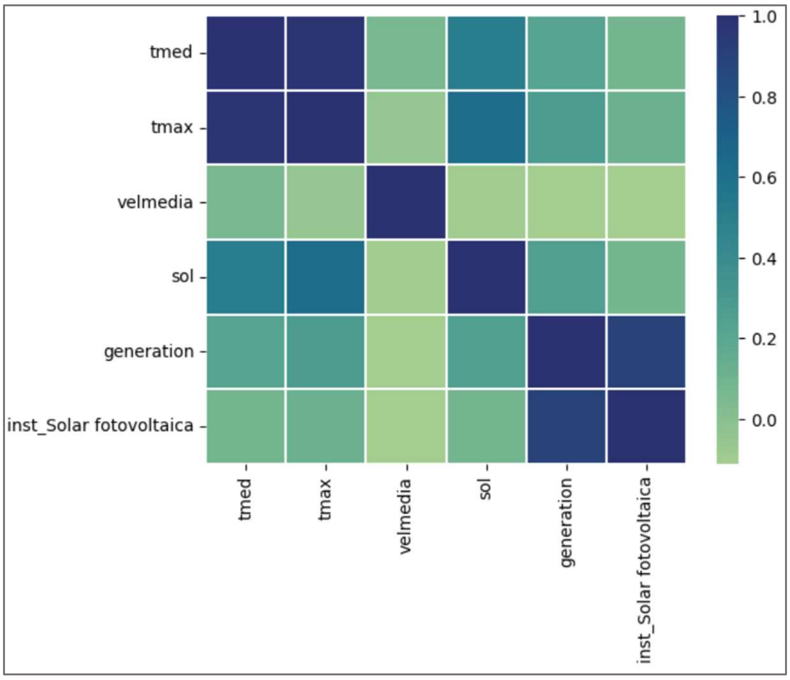
	fecha	comunidad_autonoma	CODAUTO	tmed	prec	tmin	tmax	dir	velmedia	racha	...	Eólica	Hidráulica	Solar fotovoltaica	Solar térmica	inst_Eólica	inst_Hidráulica	inst_Solar fotovoltaica	inst_Solar_térmica	demand_ccaa	Weekday
0	2015-01-01	ANDALUCIA	1	8.434	0.000	1.429	15.422	22.818	2.400	8.297	...	10798.051	1066.442	3949.598	3932.111	3325.380	588.910	875.280	1000.020	88259.789	1
1	2015-01-01	ARAGON	2	5.167	0.000	-2.033	12.383	29.417	1.818	8.428	...	9217.848	6469.745	609.142	0.000	1904.820	1338.290	168.540	0.000	22882.569	1
2	2015-01-01	CANARIAS	5	17.250	0.000	15.012	19.519	44.292	4.779	12.431	...	702.312	0.000	687.741	0.000	152.280	1.520	167.660	0.000	18441.931	1
3	2015-01-01	CANTABRIA	6	7.357	0.000	1.100	13.657	35.500	1.867	5.083	...	87.789	853.153	0.000	0.000	35.310	98.840	2.100	0.000	9493.101	1
4	2015-01-01	CASTILLA Y LEON	7	4.191	0.000	-3.507	11.926	26.646	1.653	5.680	...	17821.173	17774.025	1670.228	0.000	5556.380	4394.760	494.100	0.000	32044.289	1
5	2015-01-01	CASTILLA-LA MANCHA	8	3.808	0.000	-4.004	11.616	17.677	1.124	5.523	...	14924.135	1848.499	4539.090	1384.787	3798.930	651.110	923.420	349.400	27083.707	1
6	2015-01-01	CATALUNA	9	5.164	0.000	-0.615	10.935	40.854	1.715	6.598	...	6408.599	10451.127	825.289	51.288	1279.310	1918.640	266.430	24.290	105513.152	1
7	2015-01-01	COMUNIDAD DE MADRID	13	4.682	0.000	-2.536	11.900	22.667	0.878	5.722	...	0.000	284.384	196.497	0.000	0.000	108.520	63.360	0.000	70539.505	1
8	2015-01-01	COMUNIDAD FORAL DE NAVARRA	15	5.883	0.000	-2.383	14.150	37.200	1.600	9.900	...	4740.608	1635.210	609.142	0.000	983.200	237.700	160.610	0.000	10797.788	1
9	2015-01-01	COMUNITAT VALENCIANA	10	7.670	0.000	1.302	14.046	19.722	1.805	6.128	...	5706.287	782.057	1395.132	222.250	1193.240	641.890	347.960	49.900	60570.766	1
10	2015-01-01	EXTREMADURA	11	6.156	0.000	-1.866	14.190	20.166	1.592	5.267	...	0.000	1848.499	2534.817	2872.151	0.000	2277.360	563.860	849.000	11548.442	1
11	2015-01-01	GALICIA	12	6.403	0.000	-0.550	13.367	41.666	1.010	4.416	...	11675.941	21826.503	39.299	0.000	3309.700	3685.500	16.470	0.000	47248.766	1
12	2015-01-01	ILLES BALEARS	4	8.620	0.100	3.820	13.400	15.900	2.300	10.070	...	0.000	0.000	255.447	0.000	3.650	0.000	78.080	0.000	11403.680	1

The dataframe include all the CCAAs per day, with the weather variables for each CCAA on that date: temperature, water reservoir level, wind speed… And also information related to the powerinstalled of each technology per CCAA. 'Weekday' is also created as it can be highly related to the energy demand.

From here, the dataframe is split into smaller dataframes, one for each target. The reason behind this decision is that, in the end, each technology -or even the demand- is different to the others and is affected by weather conditions on a different way. Therefore, it was necessary to create different a different model for each target. The decision of which variables to include for each technology was based on the analysis that was made previously in the smaller notebooks and it is also based on common sense and data exploration.

For these several dataframes, several plots are painted in order to easily see how the different variables impact on the target. In this sense, pairplots(seaborn) and heatmaps(seaborn) are used, like the one on figure 7.

Figure 7: Heatmap of Solar Fotovoltaica variables, EFAT project



As weather is the main source of variables that is analyzed in the project and its impact vary so much geographically for each day, it was mandatory to have the possibility to see how the variables behaved depending on geography. It has not been possible to process all the information on a deeper perspective: by meteo station, municipalities etc. as most of the information was not provided on such an small scale. Therefore, CCAA is the administrative variable being used in the project.

In order to create a plot to draw these weather impacts by CCAA, it was needed:

-Geopandas library[2]: Open source library to work with geospatial data in python.

-GADM[3]: Open source that provides maps and spatial data for countries and subdivisions all over the world.

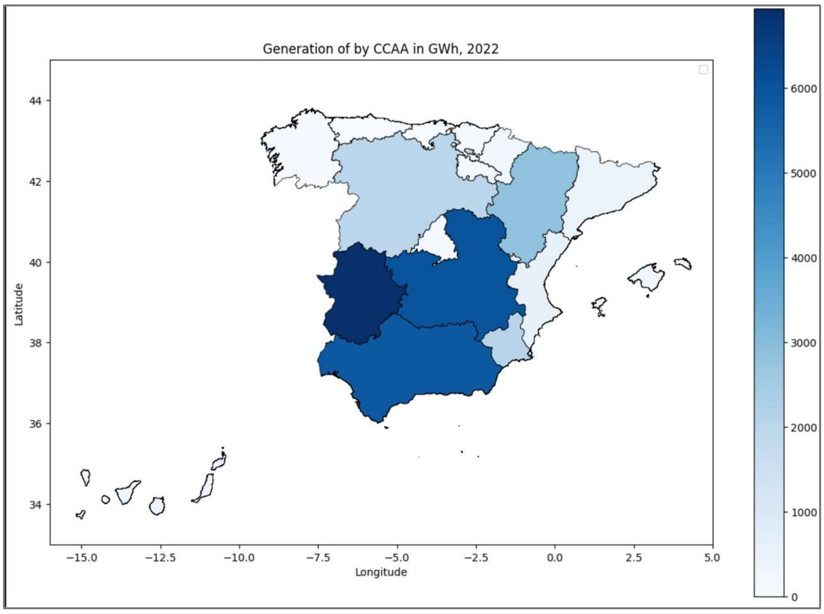
Creating a plotmap with geopandas -which works very similarly to pandas module- is easy once the basics are learnt, such as identifying the appropriate file type and understanding the downloading process. However, refining certain details, like repositioning 'Islas Canarias' to visualize it closer to the mainland, adds complexity to the task.

[2] See: GeoPandas 0.14.0 — GeoPandas 0.14.0+0.g0eb2a5e.dirty documentation

[3] See: GADM

The outcomes of this effort can be seen in the notebooks and on Streamlit. The effort is well worth it for the clarity and coherence of the map presentation.

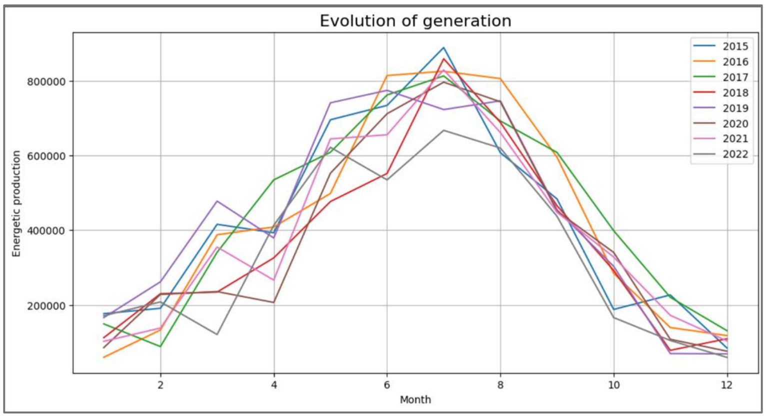
Figure 8: Solar Fotovoltaic Generation in GWh in Spain in 2022, EFAT_Project



In the "Model & Conclusions notebook", several map plots are generated for each technology to identify which CCAAs are important for each technology. From this, we can deduce that solar technology is predominantly significant in the southern part of Spain, while hydro (hidraulica) and wind (eolic) technologies are more prominent in the north. While this distribution might be anticipated, it's interesting to visually confirm it. By also mapping the weather variables, we can discern specific patterns and relationships.

Other interesting graphic to see how these technologies work, for example, is the one that shows the evolution of the generation over the year for a concrete technology:

Figure 9: Solar Thermo generation evolution over a year, EFAT_project



In this plot it is easily to see also how weather conditions that are predominant in some months of the year can affect to the target variable.

Prior to the modelling chapter, two other steps were performed too:

- **Scaling**: To ensure all the predictor variables (x-variables) were on the same scale, both StandardScaler() and MinMaxScaler() from sklearn were utilized. Both techniques were experimented with to determine which provided better results in modeling.
- **OneHotEncoder**: As CCAA has been demonstrated as a key variable to analyze, it is necessary to transform it from categorical type to binary type for a better perform of the models. OneHotEncoder from sklearn is applied to all the dataframes for this purpose.

3. Creating the models

With the dataframe now including the one-hot encoding, it's time to develop a model to evaluate the viability of the project. Sklearn offers a variety of ready-to-use algorithms. Multiple algorithms are tested across all dataframes to determine the best fit for each technology.

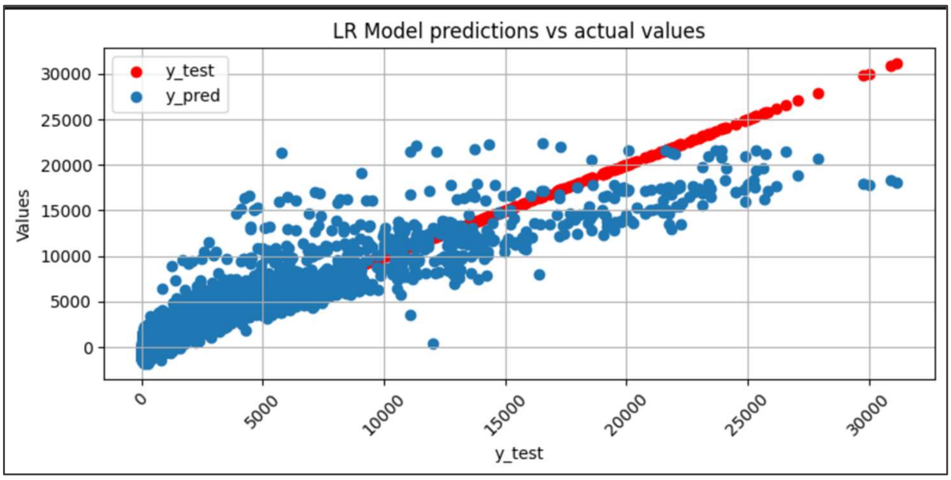
The task to solve is a regression problem: EFAT aims to predict the percentage of renewable electricity that will be generated in the upcoming days based on weather conditions. To achieve this, various regression models - falling under the category of supervised learning - available in sklearn are evaluated, from simple ones as LinearRegressor to other more complex and heavies as RandomForest.

In the ‘Predictions’ notebook the model functions already include hyperparameters tunning through CrossValidationsGrid in order to analyze, for each model, which parameters work better. It is applied to all the models.

The models tested are:

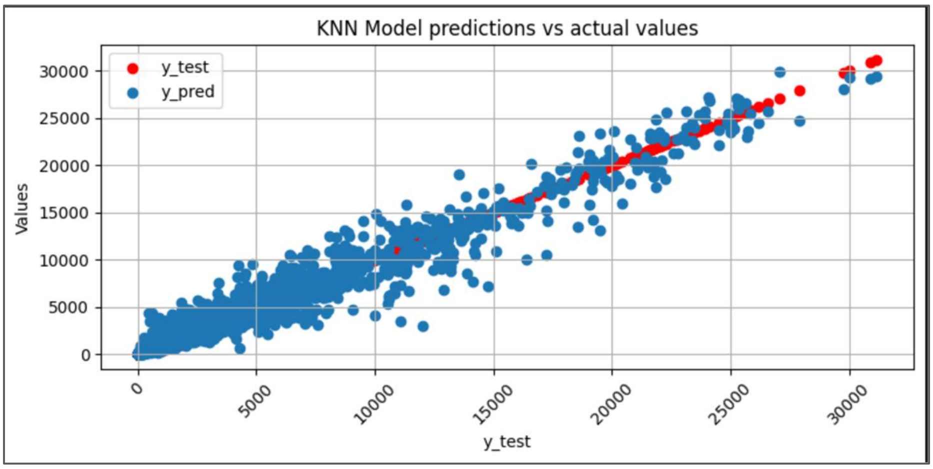
- LinearRegressor: As there are not many variables in the end for some targets and some variables seem to have lineal correlation with the target (installation capacity, for example), it was the first attempt. Results are not bad in metrics such as r2, but is worse than others.

Figure 10: Actual vs Predictions Plot of LR Model for Solar Photovoltaic electricity, EFAT_Project



- RidgeRegressor: In order to achieve a better performance in the model by avoiding multicolineality, Ridge (L2 penalty) and ElasticNet were tested. The results were not as good as it seemed, probably due to the reason that there are not many variables. Several alpha values were tested with CrossValidations methods.
- ElasticNet: Similar to Ridge, but it combines penalization L1 & L2 to the data. Even with several penalty values and adjusting the ratio between L1 & L2, the results were not promising.
- Nearest Neighbors: as it says on sklearn documentation, “KNeighborsRegressor implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user”. That k value was tested also through CVGrid. In the end, nearest neighbors brought great results in terms of r2 and is ligher than other models more complex as Random Forest.

Figure 11: Actual vs Predictions Plot of KNN Model for Solar Photovoltaic electricity, EFAT_Project



- Random Forest: As a final step, the ensembled model was tested also applying hyperparameters. It was the one that gave the best results for all the technologies. However, it is much heavier than KNN model when saving it on the cloud, which makes it less flexible and quick to be used.

Generally, the more intricate the model, the better the results it produced. Ultimately, for all the tested dataframes, Random Forest seemed to be the top choice in terms of the r2 metric. However, its performance was only marginally better than KNN.

Taking this into consideration, the model chosen for frontend implementation was KNN. A primary reason for this decision was KNN's faster download and application speed with new data. Since acquiring new data already takes considerable time, it didn't justify the use of a more resource-intensive model like Random Forest.

4. Applying the results

Once the models are calculated, it is time to apply to real data to get the information we want from weather forecasting. However, this implies a lot of work of preprocessing data again:

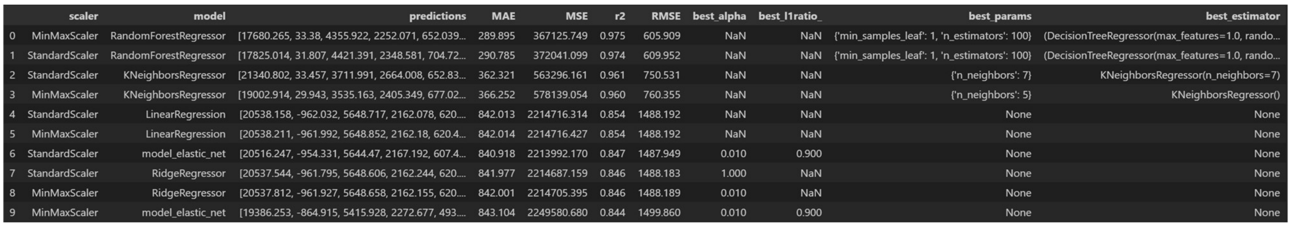
- Download new data: most recent data has to be downloaded from the same places used for historical data. The process is a little bit different though as the information is not the same:
  - Weather: AEMET does not provide an exact forecast for a CCAA or province through the API and does not provide the information in the same way for some variables such as 'Sun'. The API only allows to download forecast for next week by municipalities. This led to a tremendous amount of work to achieve a function that provided good results but at the same time was not so difficult to handle. The preprocessing of the information once downloaded is similar to what has been done with historical data: create some new features, transform the names of several CCAAs…
  - PowerInstalled: again, the API from REE is limited to extract data year by year and on a monthly basis. The function designed to download the information is also slow due to the high level of iterations over the API needed. The results are the same to what was done for historical data so the same transformations and functions are applied.
  - MiTeCo: Another complicated information to download as it is stored on a zipfile on the web and is updated every week. The zip file contains a access mdb file. The function created download the info, process it and returns a csv file. Again, it is complicated to obtain the data, which means more delay when downloading all the on-live data at the same time.

All these steps are included in the 'EFAT_Predictions' dataframe. The preprocessing and how to download the information were the main challenges, as to apply the model to the data in the end was easy to achieve with jobllib library,. Downloading the models that were saved before on the Google Drive's shared folder.

4. Results and Conclusions

As commented before, all the models prepared have been tested with all the technologies and the demand. For all the target variables, the test result was the same: the best estimators were RandomForest & NearestNeighbor. Below it can be found an example of the analysis executed with one of the technologies: 'solar foto':

Figure 12: Solar Fotovoltaic regression model clasfficated by r2 score



	scaler	model	predictions	MAE	MSE	r2	RMSE	best_alpha	best_l1ratio_	best_params	best_estimator
0	MinMaxScaler	RandomForestRegressor	[17680.265, 33.38, 4355.922, 2252.071, 652.039...	289.895	367125.749	0.975	605.909	NaN	NaN	{'min_samples_leaf': 1, 'n_estimators': 100}	(DecisionTreeRegressor(max_features=1.0, rando...
1	StandardScaler	RandomForestRegressor	[17825.014, 31.807, 4421.391, 2348.581, 704.72...	290.785	372041.099	0.974	609.952	NaN	NaN	{'min_samples_leaf': 1, 'n_estimators': 100}	(DecisionTreeRegressor(max_features=1.0, rando...
2	StandardScaler	KNeighborsRegressor	[21340.802, 33.457, 3711.991, 2664.008, 652.83...	362.321	563296.161	0.961	750.531	NaN	NaN	{'n_neighbors': 7}	KNeighborsRegressor(n_neighbors=7)
3	MinMaxScaler	KNeighborsRegressor	[19002.914, 29.943, 3535.163, 2405.349, 677.02...	366.252	578139.054	0.960	760.355	NaN	NaN	{'n_neighbors': 5}	KNeighborsRegressor()
4	StandardScaler	LinearRegression	[20538.158, -962.032, 5648.717, 2162.078, 620....	842.013	2214716.314	0.854	1488.192	NaN	NaN	None	None
5	MinMaxScaler	LinearRegression	[20538.211, -961.992, 5648.852, 2162.18, 620.4...	842.014	2214716.427	0.854	1488.192	NaN	NaN	None	None
6	StandardScaler	model_elastic_net	[20516.247, -954.331, 5644.47, 2167.192, 607.4...	840.918	2213992.170	0.847	1487.949	0.010	0.900	None	None
7	StandardScaler	RidgeRegressor	[20537.544, -961.795, 5648.606, 2162.244, 620....	841.977	2214687.159	0.846	1488.183	1.000	NaN	None	None
8	MinMaxScaler	RidgeRegressor	[20537.812, -961.927, 5648.658, 2162.155, 620....	842.001	2214705.395	0.846	1488.189	0.010	NaN	None	None
9	MinMaxScaler	model_elastic_net	[19386.253, -864.915, 5415.928, 2272.677, 493....	843.104	2249580.680	0.844	1499.860	0.010	0.900	None	None

The main metrics revised during the model evaluation were:

-**r2**: measures the model's ability to fit the target data. For all technologies in the EFAT_Project, RandomForest and KNN consistently delivered r2 values exceeding 0.8, indicating promising results.

- **MAE (Mean Absolute Error):** average of the errors between test and predicted values. In this sense, the models seem to have lower performance. Despite the figures are not extremely high, the results are not as good as in r2 terms. This disparity might be attributed to potential missing variables that could further refine the model.

The EFAT Project has been a journey riddled with challenges, surprises, and obstacles at every corner. Data extraction was already a challenge, as AEMET and REE provided great data but not always presented in the most desirable way. This necessitated extensive preprocessing: comprehending the data, cleaning it, modifying, and even formulating new features.

Yet, throughout this journey, learning has been a constant, and data was interesting enough to work with. The goal of the project (providing a tool to predict the renewable electricity production based on weather forecast) Is partially accomplished:

- The tool is created and it has consistent models behind to make it work. The results, after several days of testing, are not far from reality.

- However, it is difficult to predict weather conditions from one day to another, and that makes the tool not entirely reliable.
- Moreover, electricity production isn't solely dictated by weather. As highlighted in our documentation, various political and economic factors can influence on the matter. Especially when is about the electricity price, which was one of the firsts targets of the project.
- To try to create a model for each technology only with the information from AEMET & REE makes the project more complicated as it requires more efforts in every sense and less dedication to each technology. Specific data for each technology may be crucial to achieve better results.

On the whole, It has been an amazing journey and the results are on the table. Those results are promising and the idea of building an app that can predict electricity market trends is still alive and the project is one step ahead in that sense. But it is something that could be improved in the future: with more resources, going deeper in the geographical analysis or adding richer information related to several variables.

5. EFAT App

The way EFAT project is presented to the final user is throughout an application built on streamlit that provides access to several plots in order to understand how the model works and to have a look at historical data, but the main sheet would be the one called 'Predictions' in which the magic is done: the app downloads real-time data regarding weather forecasting for the following days in Spain and applies the pre-trained models to the data to ultimately know which percent of electricity production is going to be renewable on the upcoming days.

Link to the streamlit app:

https://efat-project.streamlit.app/

Once on the app:

1. Portait: page to explain the project and the motivation behind.
2. Data exploration: It interactively allows the user to plot several variables in order to understand the data and to see geographic trends for each variables: Map plot shows how the electricity is generated or demanded on each CCAA of Spain, Evolution over the year shows how the target variable evolves during a single year and then it is also possible to take a look at the data.
3. Predictions: The tool itself, which allows just by accessing it to download the weather forecast information and predict the % of demand that is going to be satisfied with renewable resources the following days. It works automatically, the user just have to wait for the results.