uc3m | Universidad Carlos III de Madrid

MSc in Statistics for Data Science
2020-2021

Final Master Thesis

"Estimation of Wind Energy Production in Spain"

Danyu Zhang

Supervisor:

Ricardo Aler Mur

Madrid, 2021

Abstract

Renewable energies continue to advance in Spain and in the year 2020 they have reached a new record in electricity generation with a share of 44% of all electricity production, up by 12.8% compared to the past year, according to the data from Red Eléctrica de España (REE).

Wind energy is the most mature and developed renewable energy. It generates electricity through the power of the wind, by using the kinetic energy produced by the air currents' effect. It is a clean and inexhaustible source of energy, which also reduces the emission of greenhouse gases and preserves the environment. It supplied electricity to 21.9% of Spain in 2020, namely equivalent to the electricity consumption of about 16 million households. Furthermore, wind energy represents 0.35% of Spanish GDP.

In light of the openness and transparency of the *ESIOS* database (Sistema de Información del Operador del Sistema) that provides the electricity production and *ECMWF* (European Centre for Medium-Range Weather Forecasts) which provides the meteorological information through their own APIs, the purpose of the project is to estimate both the wind energy production of the province of Cadiz and the entire country, maintaining the smallest possible error margins while modelling. We utilize meteorological variables in the target region's corresponding coordinates (both forecasted values and 'real' meteorological values), so that this proportion of the energy production can be planned into the national electrical system.

During the model building process, various supervised machine learning methods will be employed: *Random Forest, Gradient Boosting* and *Support Vector Machines*. Furthermore, due to the immensity of the data sets, 2 dimension reduction methods will be applied, *Principal Component Analysis* and *Partial Least Squares* in order to speed up the algorithms. After hyper-parameter tuning, the model that has the best metrics is *SVM* for both regions following the application of *Partial Least Squares* techniques.

Keywords: Supervised Machine Learning, Regression, Hyperparameter Tuning, Evaluation, Wind Power

Content

Chapter 1

Introduction and objectives

1.1. Introduction

The development of renewable energy sources is one of the key aspects of the Spanish national energy policy. They efficiently contribute to the reduction of greenhouse gas emissions (like CO2), reduce our dependence on petroleum products and diversify our sources of energy production by promoting renewable resources such as solar, wind, etc.

The renewable energy production of 2020 in Spain has hit the historical maximum of renewable coverage during the past 10 years: 44%. Additionally, the evolution of the share of renewable coverage is increasing yearly, from 31% in 2011 to 44% in 2020.

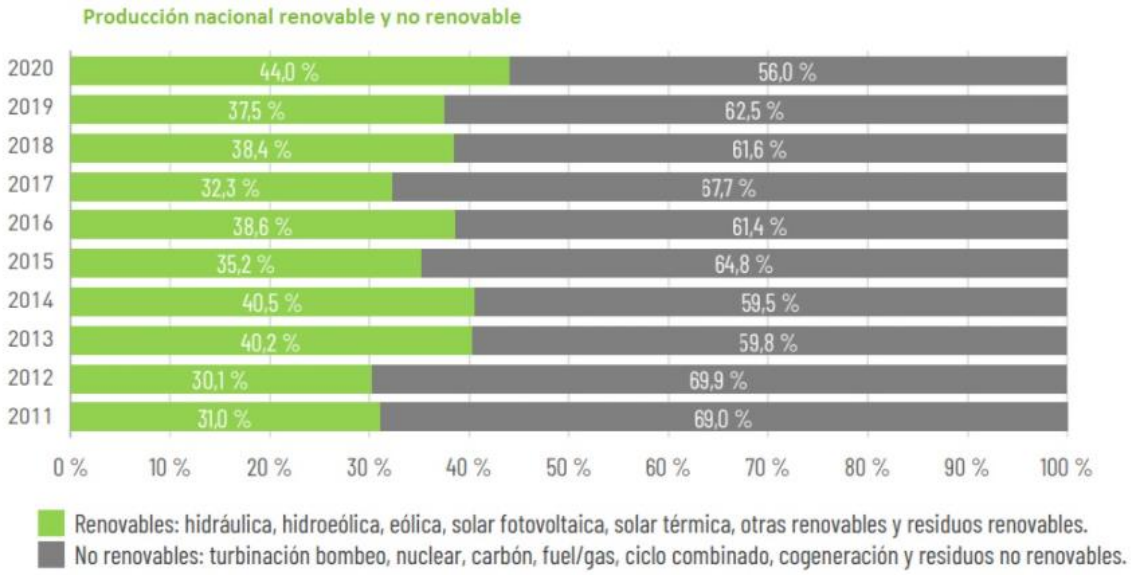

Figure 1.1 Spanish national production of renewable and non-renewable energy during 2011-2020 (Provided by *Energía Renovables*: https://www.energias-renovables.com/panorama/espana-ha-producido-en-2020-mas-electricidad-20210312)

In this project, the focus is on the estimation of wind energy production. It is a popular sustainable, clean, renewable energy source that has very little impact on the environment.

Within the context of the Spanish electricity supply market, since the first aerogenerator was started up in Catalunya on March 10, 1984, until 2020 the wind has produced 21.9% of the national electricity demand (almost 50% of renewable coverage), it has become the most efficient way of renewable energy production.

Nevertheless, unlike with non-renewable energy, the variability is one of the main limitations of renewable energy resources. Wind electricity production depends on the meteorological conditions at every instant. It is a reason to plan how this share of energy production should be consumed to estimate or forecast the electricity production of what is seemingly an unmanageable resource.

This project will use diverse machine learning methods from the *Python* package *sklearn* to predict the wind energy produced on different time horizons, 8 time horizons per day in 3-hour intervals. Concretely, the methods are ***Random Forests, Gradient Boosting*** and ***Support Vector Machines***.

The modeling will be regional, estimations of electricity production for Cadiz and Spain will be done. Unlike ordinary modeling problems, this problem is addressed by defining a spatial grid of points (coordinates) per region, where *ECMWF* (European Centre for Medium-Range Weather Forecasts) provides the meteorological variables as the predictors for each point of the spatial grid. This leads to a consequence that, the dimension (quantity of variables) of the data set could be very large depending on the region area: for the entire country there are 3536 variables. Therefore, dimensionality reduction techniques need to be applied before modeling with machine learning methods. ***Principal Component Analysis (PCA)*** (unsupervised) and ***Partial Least Squares (PLS)*** (supervised) will be used. Despite *PLS* is not so well known as *PCA*, the motivation to use it is that it takes into account the target variable, and therefore, it may work better in supervised regression problems, as the one addressed here.

For Cadiz, both predictors' set: **forecast** (*ECMWF* computes an ensemble (set) of predictions with 12 ensemble members, the final prediction is their average) **of meteorological variables** and **'real' meteorological variables** (in fact they are also estimations, but they are as close as possible to actual measurements) provided by *ECMWF* will be trained in order to check that, if the fact that

in ‘real’ meteorological variables there is less uncertainty in the values favours to improve the models.

Furthermore, we also want to study the importance of feature engineering (adding more variables that are the wind speed modules per each coordinate). Since the energy production depends on the amount of wind, does including these new variables help?

Lastly, the machine learning interpretation methods such as *Permutation Test, Partial Dependence Plot* will be utilized not only for the purpose of understanding the relationship between the meteorological condition and amount of electricity production, but as well to validate the models.

1.2. Objectives

The primary objective of the project is to estimate the electricity production of Spain within the smallest possible error margins employing machine learning techniques. Based on this main objective, the following partial objectives are proposed:

- Querying of the data bases: electricity production (*ESIOS* database) and meteorological information (*ECMWF* database) using their APIs.
- Pre-processing: data transformation and data merging.
- Understanding the relationship between the meteorological conditions and the electricity production.
- Modeling and hyper-parameter tuning using machine learning tools in *Python*.
- Interpretation of the models to obtain conclusions.
- Model evaluation through the use of different metrics. In particular, in addition to determine which machine learning method works better for this problem, we intend to evaluate whether dimensionality reduction techniques (*PCA* and *PLS*) and feature engineering techniques, are advantageous for this problem. This will be done for two regional models: a small one (Cadiz) and a very large one (Spain).

1.3. Memory Structure

The structure that specifies the memory is the following:

- **Data:** Explanation of data sets: explanation of meanings of each variable, the different time horizons, train-test splitting, etc.

- **Machine Learning Methods:** Introduction to basics of machine learning discipline.
- **Energy Forecasting for Cadiz:** Selection of the best model for Cadiz using specific metrics such as explained variance, rooted mean square error, study the relationship between features and the response in order to facilitate the analysis of the country.
- **Energy Forecasting for Spain:** Selection of the best possible model for Spain.
- **Conclusions:** The conclusions of the project will be drawn, which are the most important facts for the estimation of the energy, how it would be theoretically possible to further improve the model, etc.
- **Appendix:** *GitHub* repository URL and QR code (containing the URL).

Chapter 2

Data

Data is the essential element of a project, it plays a very important role in the analysis. For each region, 3 data bases during years 2015-2018 are going to be used: the **forecasts of meteorological variables** (8 time horizons per day), **'real' values of meteorological variables** (24 time horizons per day) and the **response variable**: electricity production (24 time horizons per day). Last two databases need to be transformed into 8 time horizons per day.

2.1. Sources

The data used during the project are from 2 different sources. On the one hand, the electricity productions (response variable) are collected from *ESIOS* (Sistema de Información del Operador del Sistema); on the other hand, the meteorological variables (the predictors) are collected from *ECMWF* (European Centre for Medium-Range Weather Forecasts).

- ***ESIOS***: The data can be obtained from the website by clicking analyze indicator in generation and consumption, selecting period, group, compare, and finally, export data. The format can be selected between CSV, JSON and EXCEL.

Figure 1.2 Home page of ESIOS (https://www.esios.ree.es/es).

- ***ECMWF***: To extract meteorological information, the first thing needed is to create an account in https://cds.climate.copernicus.eu/ , then specify the requirements filling the following form including the coordinates that include the region interested, range of time, variables selected, etc: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form.

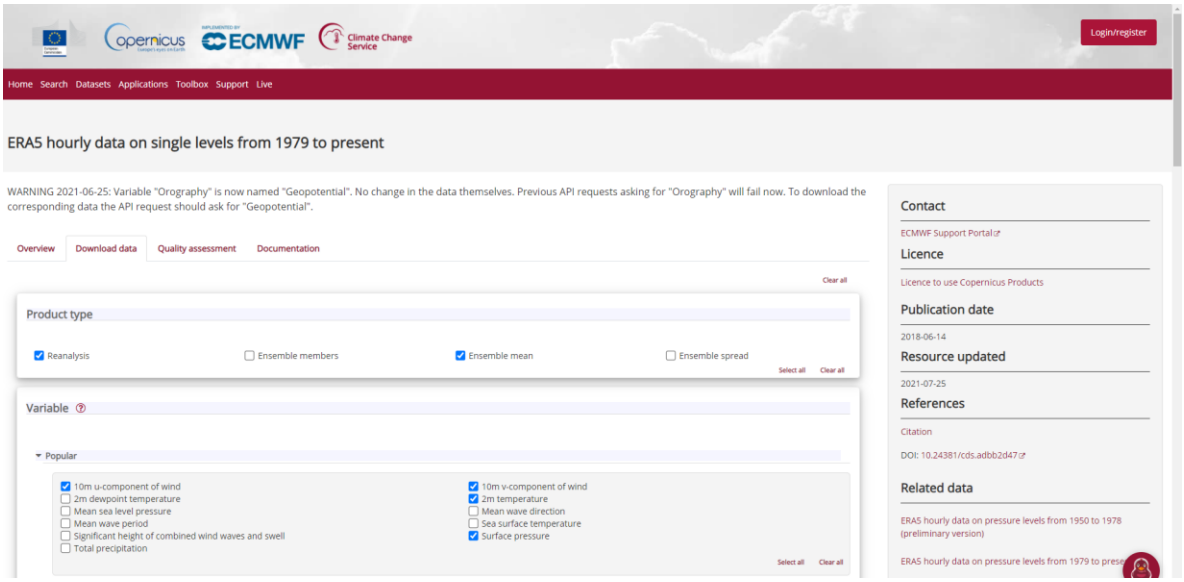

Figure 1.3 An example of filled form of ECMWF.

2.2. Description of Data

ECMWF provides the meteorological variables (the predictors) distributed on a spatial grid. For each point in the grid, *ECMWF* is able to provide both forecast variables and 'real' variables. The **forecast variables** are meteorological predictions carried out by a mathematical model. They are also called '*ensemble mean*' because *ECMWF* computes an ensemble (set) of predictions with 12 ensemble members, and the final prediction is their average. The **'real' variables**, also called *reanalysis*, they are not measurements but also estimations done by a model as close as possible to the true measurements. They help to remove the uncertainties related to meteorological forecasts. Forecast variables are provided every day for 8 horizon predictions during the day (that is, one forecast every three hours). *Reanalysis* ('real') variables are provided every hour.

For the meteorological variables (predictors): both forecasts done by *ECMWF* and the 'real' (reanalysis) meteorological values, 6 different variables will be extracted for each grid point of

each region (coordinates on the map), and 2 more variables will be added by using feature engineering, all of them are numerical:

- **10m u-component of wind**: It is the horizontal speed of air moving towards the east, at a height of 10 metres above the surface of the Earth, in metres per second.
- **10m v-component of wind**: It is the horizontal speed of air moving towards the north, at a height of 10 metres above the surface of the Earth, in metres per second.
- **100m u-component of wind**: It is the horizontal speed of air moving towards the east, at a height of 100 metres above the surface of the Earth, in metres per second.
- **100m v-component of wind**: It is the horizontal speed of air moving towards the north, at a height of 100 metres above the surface of the Earth, in metres per second.
- **2m temperature**: It is the temperature of air at 2m above the surface of land, sea or in-land waters.
- **Surface pressure**: It is the pressure (force per unit area) of the atmosphere on the surface of land, sea and in-land water.
- **Norm (modulus) of the velocity at 10 meters**: Calculated by $\sqrt{u\_10^2 + v\_10^2}$
- **Norm (modulus) of the velocity at 100 meters**: Calculated by $\sqrt{u\_100^2 + v\_100^2}$

Furthermore, it is worth mentioning that, for each region there will be:

$$(\text{Nº of coord. on Longitude} \times \text{Nº of coord. on Latitude} \times 8 \text{ Meteorological variales}) \text{ Predictors}$$

For example, in Cadiz, there are in total 12 grid points, so there are $12 \times 8 = 96$ predictors.

For the response, the only variable that will be used is the value of the energy production:

- **Value**: Value of the energy production in real time (obtained from *ESIOS*).

There are no missing values nor null in all the databases.

2.3. Preprocessing: Transformation and Data Merging

Firstly, as the predictors' data bases extracted from *ECMWF* have *NetCDF* (Network Common Data Form) format, it is necessary to transform them into another format that can be treated in *Python*. To solve this problem, a loop is used to go through every row of the data bases.

The second step is to merge the predictors with the response into one single data set, in order to do so, it is required to have the same time format to match them.

- **Merging forecasts meteorological variables with the response variable:**
  The *ECMWF* forecasts have information of only 8 time horizons per day available, as we have hourly data for electricity production, it's necessary to sum it for each 3 hours in order to fit the models.
  Also, in this data set, there are 4 repeated times for the response variable, which is caused by the changes between standard time and daylight-saving time. The decision made is to be pessimistic about estimating the electricity production, it's always better to have more than needed, so it is reasonable to take the smaller value.
- **Merging 'real' values meteorological variables with the response variable:**
  In this case, the meteorological variables match the electricity production (both databases from *ECMWF* and *ESIOS* provide their variables every hour). The idea of using 'real'/*reanalysis* variables is to use them to train the models, then check whether non-noisy variables help to obtain better results. However, the model will be eventually used for forecasts predictors, because at that moment, *reanalysis* variables will not be available (*reanalysis* variables are only available for the past). Therefore, only *ECMWF* forecast variables can be used as predictors at testing time. For this reason, we will take the means for each 3 hour of the meteorological values as the new predictors, and the sum for each 3 hour of the energy production as the new responses, so that training and testing data match, time-wise.

In addition, for the country modeling, since the response data bases are regional, it is required to match the time columns for each region and then sum the productions to obtain the national total.

After this procedure, we have **2 data sets for Cadiz** which have **distinct predictors** (forecasts and 'real' values of meteorological variables) and the **same responses** (electricity production). The purpose is to identify which model trained from different data sets works better for further estimation of electricity production in Spain.

Data from **2015-2017** is used as the training set, where the data of **2017** is going to be used as the inner validation set (for hyperparameters tuning); data of year **2018** is used as the testing set (outer validation set) where the metrics' results are obtained with the intention of comparing between

diverse models. The reason why the data can't be shuffled randomly to do train-test splitting is that, since it is a time series database, we can only use the previous data to predict the later data.

All the data extraction, transformation and data merging codes used are updated on *GitHub*, which can be found on the last chapter, appendix.

2.4. Dimension Reduction: PCA and PLS

The amount of variables depends on the area of each region one wants to estimate, namely, there are 8 variables per each coordinate (grid points) on the map.

To be clear, there are **96 variables for Cadiz** and **3536 variables for Spain**. Which are quite considerable quantities for the modeling, thus, reducing the dimension is a one of the solutions to become more efficient. *Principal Component Analysis* and *Partial Least Squares* are effective ways to solve the problem.

2.4.1. Principal Component Analysis

PCA is used to transform linearly the original data set to a new data matrix such that, the first principal component (new variable) contains the most relevant information of the original set, which means that, it is the linear combination of the original variables with the largest sample variance; the second principal component contains the most relevant information in the original set that is not included in the first PC; and so on. Thus, we can focus on analyzing the first variables of the new data matrix instead of the original ones. The new variables (principal components) are uncorrelated. Furthermore, it is fundamental to scale the original features since it is a variance maximizing method.

In *PCA*, the transformation is unsupervised, meaning that no information about the response variable is used.

The hyper-parameter that is tuned in *PCA* is the percentage of variance that the method explains. This determines how many components are selected.

2.4.2. Partial Least Squares

It finds a linear regression model by projecting the response and the predictors to a new space (projection to latent structure). It is quite similar to *PCA*, which also applies a dimensionality

reduction to the variables. The main difference between them is that *PLS* transformation is supervised, it considers the response information (of training set). Therefore, it might be more useful for supervised problems, such as the regression problem solving during this work.

For *PLS*, we will transform the amount of variables into different numbers according to the optimization done for both Cadiz and Spain.

Chapter 3

Machine Learning Methods

Machine learning is a process of building inductive models that learn from a limited amount of data without human's intervention. In order to solve our supervised regression problem, 3 machine learning methods are used for the estimation: *Random Fores*t, *Gradient Boosting* and *Support Vector Machines*.

Furthermore, the technique of hyper-parameter tuning will be employed in each of the machine learning methods. The procedure is, firstly, to define the search space for every single hyper-parameter; through *random search*, some combinations of hyper-parameters are randomly tested, the prediction error is calculated with the validation data (year 2017). Lastly, the metrics (by using data from year 2018) of all the models are compared, the best model is obtained with the combination of hyper-parameters that has the lowest prediction error.

1. Random Forest

Ensemble learning is a model building technique that is typically more accurate than base models such as *linear regression*, *decision trees*, etc. It creates a final model based on a collection of individual models, classifying data by voting and predicting by their mean.

Random Forest is one of the most popular ensemble learners, a type of *Bagging*. It builds decision trees in **parallel**; each tree is built with different training sets (random observations taken with replacement), the average of every tree is taken for the final prediction. As it utilizes the resampling method bootstrap, this process introduces randomness to the tree generation, in such a way that it will reduce the variance of the data set, avoiding overfitting.

The hyper-parameters that will be considered during the project are as follows:

- ***max_depth:*** the maximum number of levels in each decision tree (for every randomly created samples), a large value can cause overfitting to the training set;
- ***min_samples_leaf:*** the minimum number of instances to grow a new leaf node, a small value can introduce overfitting as the tree specifies too much the training set;
- ***n_estimators:*** the number of trees in the forest; number of samples one wants to create. As it increases, the robustness will be higher with more computer cost;
- ***max_features:*** the maximum number of features considered for splitting a node, it is not recommended to use all of them not only because it takes relatively long time to run, but, it causes overfitting.

2. Gradient Boosting

Gradient Boosting is very widely used as well. Unlike *bagging* methods that builds base models parallel, *boosting* builds them **sequentially**, in such a way that each base model is dedicated to solve the difficulty of the previous one. The idea of *Boosting* is to improve weak models. Base models are added to the set consecutively, in such a way that the next model tries to understand the errors of the previous model.

Back to *Gradient Boosting*, the first prediction will be the mean of the response variable, then the next model is constructed using the residuals of the previous model as new response variable in each iteration, some coefficient will be put which measures the importance of the pre-model. The final model will be the collection of every single model of the ensemble.

The hyper-parameters that are considered are:

- ***learning_rate****:* it determines the weight of each tree added to the global model;
- ***n_estimators****:* the number of trees in the forest, the larger the value is, the better the result will be until it hits the optimal;
- ***max_depth****:* the maximum number of levels in each decision tree in the forest, small value will cause underfitting while large value causes overfitting;
- ***min_samples_leaf****:* the minimum number of instances to grow a new leaf node;

3. Support Vector Machines

The method *Support Vector Machines* was originally created to solve classification problems, but nowadays, it is used very commonly for regression problems as well. *Support Vector Machines* for Regression (*SVR*) uses the same principle as *SVM* for classification. *Support vector machines* for classification (*SVC*) transform the data set into a higher dimension where the data can be separated using the transformation function. When applying it to regression, the objective is to learn a regression model that approximates the response variable as close as possible. Furthermore, another *epsilon* parameter is added, which is used so that the learned function does not deviate from the real output by an outlier that is greater than *epsilon*.

The hyper-parameters of *SVR* that are tuned are:

- ***kernel:*** it is the transformation function used so that the data points can become linearly separable. The function will be set to *RBF (radial basis function)* to reduce the computer cost;
- ***C:*** it adds a penalty for each incorrect data point, if it is set to be large, the model will overfit the training set since the minimization of the training error has a lot of weight;
- ***gamma:*** it is a hyper-parameter of *RBF* that controls the distance of influence of a single training point, therefore, models with large gamma values tend to overfit.

Chapter 4

Energy Forecasting for Cadiz

Before estimating the electricity production of Spain, we start with the province Cadiz. It is clear to observe from the following graph that, due to its geographic location, the province possesses great wind potential (close to the Atlantic Ocean and the Mediterranean Sea). Moreover, it has just inaugurated a new wind farm in February 2020.

In this chapter, all the models' results of the province are shown, for both training set and testing set with the purpose of detecting overfitting and underfitting problems (for all the following tables, TR means "training" and TE means "testing").

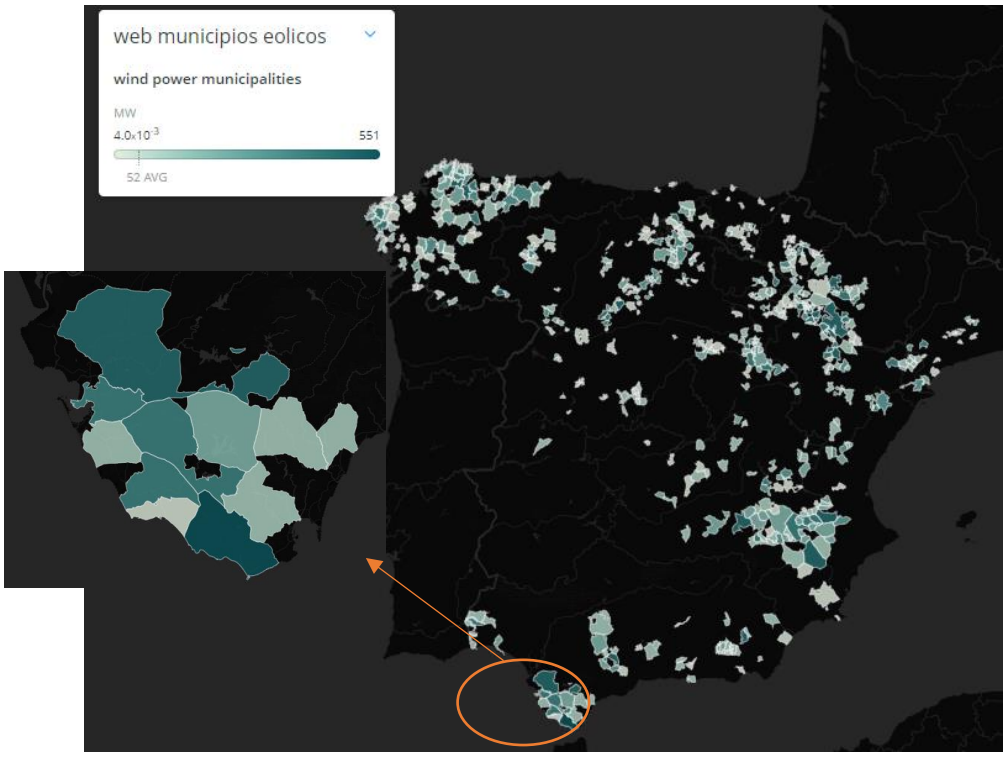

Figure 4.1 Wind Installations by town map (Source: ESIOS https://www.esios.ree.es/en/interesting-maps/wind-installations-town-map)

The metrics used to evaluate the models are as follows:

- **Explained Variance**: also known as R squared or coefficient of determination:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- **MAE**: mean absolute error, measures the average absolute errors between paired observations:

$$MAE = \frac{\sum_{i=1}^{N}\left|\widehat{Y}_i - Y_i\right|}{n}$$

- **MSE**: mean squared error, measures the average squared differences between paired observations:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2$$

- **RMSE**: root mean squared error, root of MSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}{n}} = \sqrt{MSE}$$

- **Maximum error**: estimates the maximum residual error between paired observations:

$$MAX_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|$$

Multiple linear regressions will serve as the baseline models for both forecasts and 'real' meteorological models, further advanced models are compared with it.

In order to facilitate the modeling of the Spanish electricity production, we will modify Cadiz data to **3 different sets** in order to check with which one we get the best results: training with all the predictors (section 1); training without module variables (new variables created, section 2) and training without u,v component variables (section 3).

4.1. Models trained by forecasts

4.1.1. Models trained by all variables

We can observe from the following table (Table 4.1) that, training using all the variables (96) after hyper-parameter tuning, the best results are obtained using method *Gradient Boosting* with as well the highest computational cost (**2979 seconds**, 50 minutes approximately). With almost **91% of**

explained variance, and **182 units of mean absolute error**. Comparing to the baseline model, it has improved around 3.6% of R squared and 57 units of mean absolute error.

Moreover, we can observe that in this case, the method *Support Vector Machines* did not perform well, that is because it needs a very high computer cost in order to have a higher precision, due to the availability of our resources, it has not met our expectations.

	LM	**RF (424 sec)**	**GB (2979 sec)**	**SVM (402 sec)**
R^2 TR	89.3%	94%	98.1%	96.5%
R^2 TE	**87.3%**	89.9%	**90.9%**	80.5%
Max error TR	3241.2	1553.8	741.3	1724.4
Max error TE	1970.6	1632.8	1612.5	1863
MAE TR	238.8	177.6	100.9	91.9
MAE TE	**238.5**	201.8	**181.9**	272.7
RMSE TR	329.9	246.7	137.5	188.2
RMSE TE	**327**	292.6	**276.6**	407.2

Table 4.1 The metric outputs using all the forecasts variables in Cadiz.

4.1.2. Models trained by all variables except *modules* (new variables created)

Removing the 24 module variables (*uv10* and *uv100* per 12 coordinates of Cadiz), 72 variables are left. The results have worsened little (0.3% of R squared and 7 units of mean absolute error). The smallest errors are still obtained from the method *Gradient Boosting*, it works much better than the linear model, meaning that it could be an appropriate choice if the budget is low.

	LM	**RF**	**GB**	**SVM**
R^2 TR	77.3%	93.8%	98.1%	93.5%
R^2 TE	**65.8%**	88.3%	**90.5%**	82.3%
Max error TR	3353.9	1501.1	879.1	1972.6
Max error TE	2948.7	1655.3	1668.3	2551.6
MAE TR	358.9	188.5	103.2	154.3
MAE TE	**388.5**	226.1	**189**	261.7
RMSE TR	480.2	251	139	257.7

RMSE TE	**537.6**	314.8	**283.4**	387.4

Table 4.2 The metric outputs using all minus *module* forecasts variables in Cadiz

4.1.3. Models trained by all variables except u,v component variables

To train the model without the 48 *u, v components* variables (*u10, v10, u100* and *v100* per 12 coordinates), the metrics obtained are not ideal compared to before. Hence it is necessary to keep them if the objective is to get the smallest errors possible, although training with only 48 variables makes the algorithms a lot faster.

	LM	**RF**	**GB**	**SVM**
$\boldsymbol{R^2}$ TR	86.9%	93.2%	97%	93.1%
$\boldsymbol{R^2}$ TE	**85.2%**	86.7%	**88.4%**	80.2%
Max error TR	3792.3	1677.7	1085.1	1948.6
Max error TE	2252.7	1894.5	1921.2	2077
MAE TR	267.9	188.1	126	151.8
MAE TE	**271**	229.1	**209.7**	278.1
RMSE TR	364.9	262.3	175.9	265.6
RMSE TE	**353.9**	335.6	**313.2**	410.7

Table 4.3 The metric outputs using all minus *u,v component* forecasts variables in Cadiz

To conclude, the best metrics obtained for Cadiz are from training the whole data set with *Gradient Boosting*. We will analyze which are the most weighted features in this model in order to illustrate what variables influence the electricity production of Cadiz the most.

- **Permutation Test**

The permutation test is a procedure to determine which the most important predictors in the model are. It works by shuffling the values of each predictor in turn randomly, and then testing the models with the shuffled dataset. If the predictor is important, error after shuffling will increase. Thus, this procedure allows to display the most relevant variables for the prediction.

Table 4.4 displays the results of the permutation test. Since the first number in each row shows how much the model performance decreased with a random shuffling, the most important variables according to *Gradient Boosting* is *the* ***module of wind speed in coordinate (36.0, -6.0)***, it affects

a lot to the final prediction; then ***the wind speed moving towards the east, at 100 metres above the surface in coordinate (36, -5)****;* also, from the first six most important variables, four of them are modules of speed from different coordinates, which means that, feature engineering (creation of module variables) worked practically for the modeling. The most relevant coordinates appear to be clustered close to each other, and this is probably due to important wind farms located at those positions.

Weight	*Feature*
0.0673 ± 0.0035	Module of 100 meters on (36.0, -6)
0.03313 ± 0.0014	U 100 meters component on (36.0, -5.0)
0.0265 ± 0.0032	Module of 100 meters on (36.5, -6.0)
0.0182 ± 00.0015	Module of 10 meters on (36.5, -6.0)
0.0128 ± 0.0018	Module of 10 meters on (36.0, -6.0)
0.0098 ± 0.0007	Module of 100 meters on (36.0, -5.5)
0.0075 ± 0.0010	U 10 meters component on (36.0, -5.0)

Table 4.4 Output of Permutation Test of *Gradient Boosting* using all forecast variables for Cadiz

- **Partial Dependence Plot**

PDP shows by plot how the predicted value changes if a single variable in the data set is changed, the *y* axis is interpreted as the change in the prediction from what it would be predicted at the baseline (the original prediction). We will analyze the plots of the 2 most important features according to the permutation test and the interaction between them.

For the feature ***module in coordinate (36.0, -6),*** as it increases, the energy production increases until a certain point around 12, faster speed beyond that appears to have very little impact on the predictions. The reason why when the wind speed gets extremely high, the production does not grow as fast as previously is because, if the wind is very strong, the wind turbine stops by itself to avoid possible failures or breakdowns. The fact that the variation of energy production with respect to the module follows something close to a sigmoid curve (Ahmed, 2013, https://www.researchgate.net/publication/334729294_An_Analytical_Study_for_Establishment_of_Wind_Farms_in_Palestine_to_Reach_the_Optimum_Electrical_Energy), helps to validate the model, as it is known that individual aerogenerators operate according to such curves (and regional models of wind farms should also follow them to some extent).

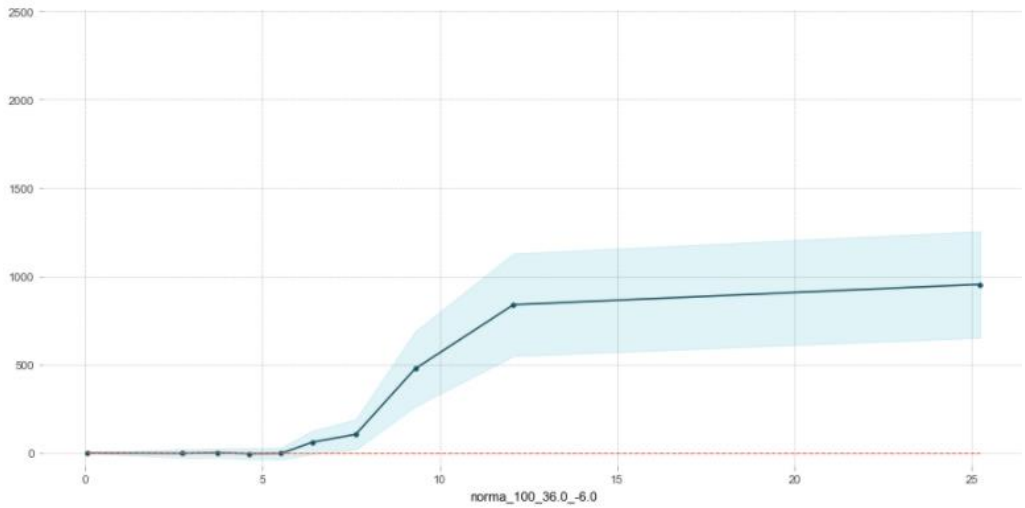

Figure 4.2 Output of ***module on coordinate (36.0, -6)*** PDP

Negative values of u components in **coordinate (36.0, -5.0)** increases Cadiz's energy production, which implies that the air moving toward the east disfavours the energy production while the air moving toward the west favours the energy production.

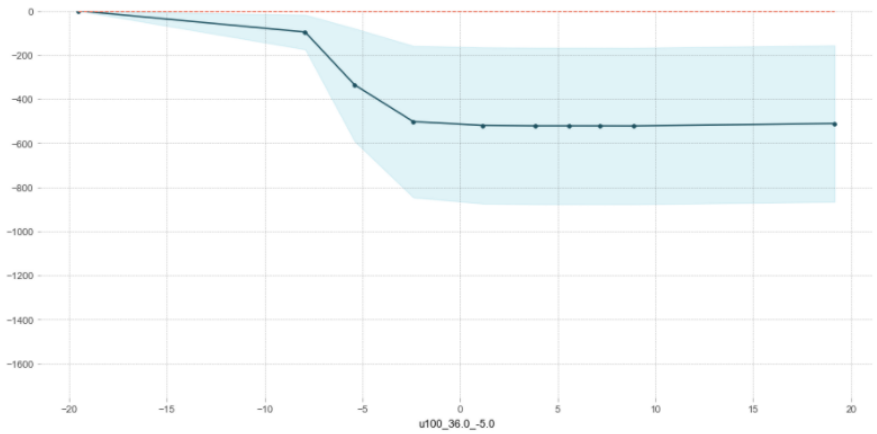

Figure 4.3 Output of ***module on coordinate (36.0, -6)*** PDP

The interaction PDP between 2 previous variables shows exactly the same conclusions obtained before. The energy production hits the peak while having the combination of lowest ***u component in coordinate (36.0, -5.0)*** with the highest value of ***module in coordinate (36.0, -6)*** (yellow area), and it is less while there are higher values of ***u component on (36.0, -5.0)*** with lower values of the variable ***module on (36.0, -6)*** (purple area).

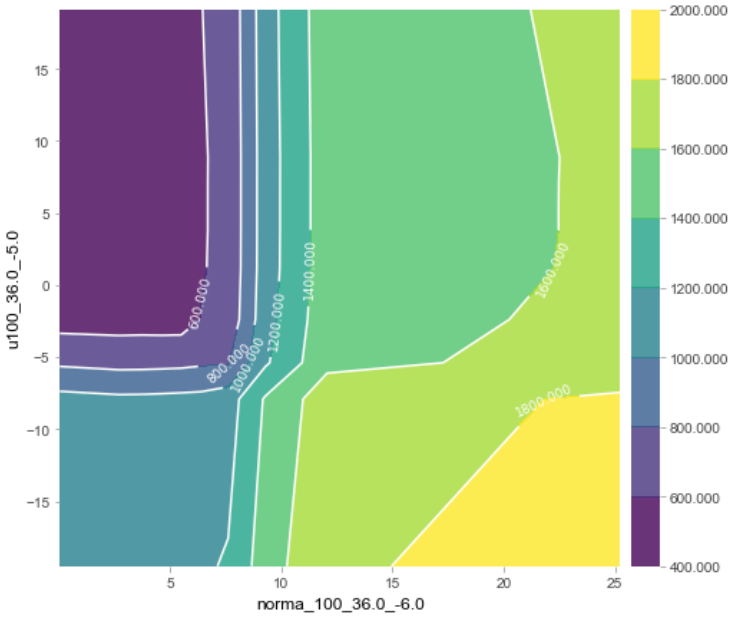

Figure 4.4 Output of interaction Partial Dependence Plot for Cadiz

- **Summary Plots**

This plot summarizes permutation importance and partial dependence plot. When larger value of ***module in (36.0, -6),*** more elevated the energy production (according to the line on the right, the color of the points is warmer); for ***u component in (36.0, -5.0),*** lower value cause the increment of electricity production in Cadiz (points with colder color tend to have higher SHAP value).

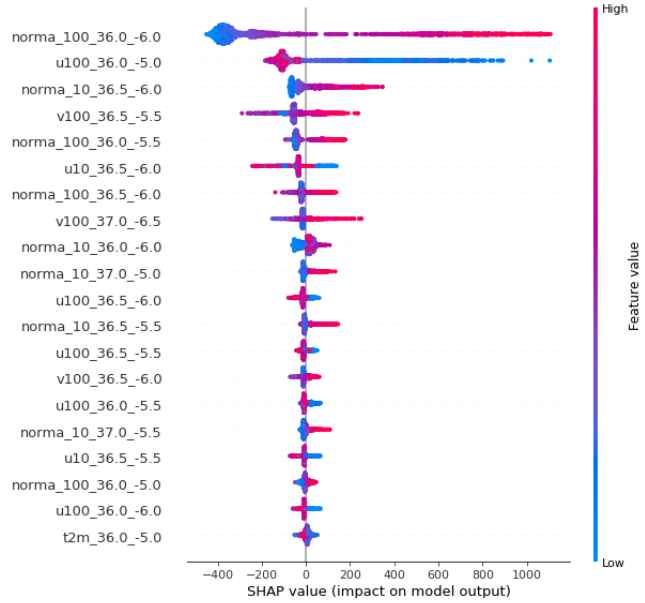

Figure 4.6 Output of Summary Plot

4.1.4. Applying Dimension Reduction: *PCA* & *PLS*

After achieving the results before, *PCA* and *PLS* will be applied to the sets.

- **Principal Component Analysis**

Table 4.5 displays the results of PCA. The best model according to *PCA* is *Gradient Boosting*. After hyper-parameter tuning, we got the best results with **95% variance explained *PCA***, which means that the model is done using only **8 new variables** (components).

PCA	**RF**	**GB**	**SVM**
R^2 TR	88.4%	95.3%	90.3%
R^2 TE	84%	**89%**	87.3%
Max error TR	1900.3	1541	2242.1
Max error TE	1971	1993.8	1870.9
MAE TR	252.5	152.1	212.1
MAE TE	263.2	**206.2**	220.8
RMSE TR	342.7	218.3	314.118
RMSE TE	367,4	**304.4**	324.995

Table 4.5 The metric outputs using *PCA* of Cadiz

The following plots shows the explained variance per each component and the cumulative explained variance:

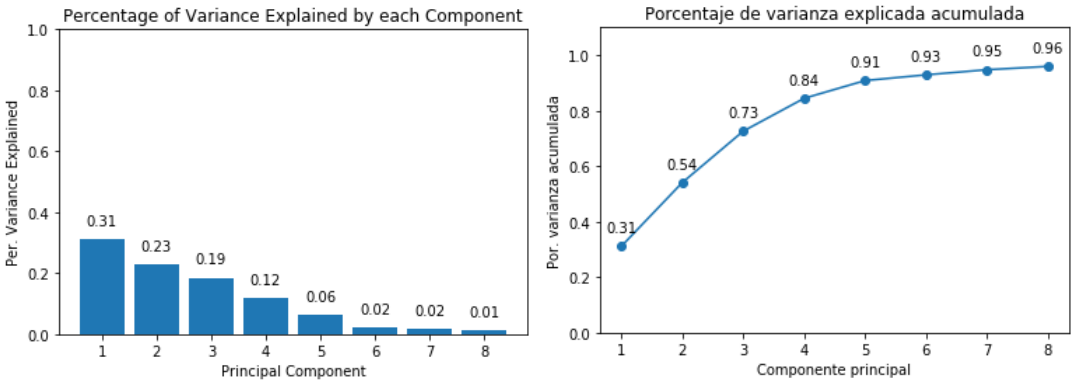

Figure 4.8 Output of Explained Variance of *PCA* of Cadiz

Although results with PCA are worse than the original results in Table 4.1., it allows to obtain reasonable results with only 8 variables.

- **Partial Least Squares**

The following graphs shows the *MSEs* and *R squareds* by using from 1 to 30 components on *PLS regression*; it is the optimal while using 21 variables for *PLS regression* (10 and 15 are also tried for the machine learning methods with the intention of avoiding overfitting, although the **best results are still obtained with 21 variables**).

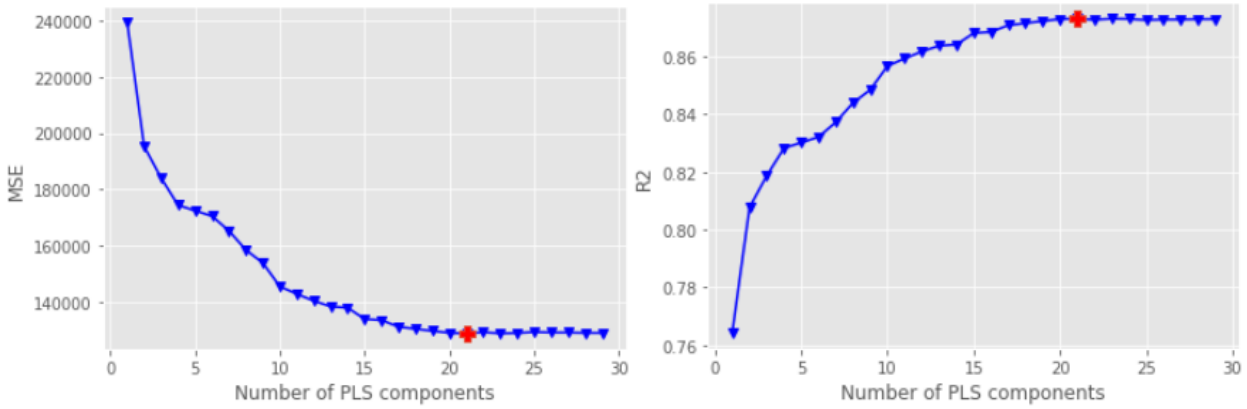

Figure 4.9 Output of Optimization for selecting the number of components for Cadiz

From the table below we can see that, by using the *PLS* method, the machine learner ***Support Vector Machines*** **has achieved the smallest error margins by far with 180 of MAE** and **91.2% of explained variance**. Although it has only improved very little from *Gradient Boosting*, even so very significant. The method has not only carried out the best results, but it has also accelerated the algorithms' speed by reducing the dimension to just 21 variables.

PLS	**RF**	**GB**	**SVM**
$\boldsymbol{R^2}$ TR	88.2%	95.7%	94%
$\boldsymbol{R^2}$ TE	84%	90.5%	**91.2%**
Max error TR	2022.6	1330.7	1824.4
Max error TE	1965	1530.4	1565.8
MAE TR	251.1	147.9	166.4
MAE TE	257.8	191.971	**179.9**
RMSE TR	346.3	209.3	248
RMSE TE	367,8	283.7	**273.5**

Table 4.6 The metric outputs using *PLS* of Cadiz

4.2. Models trained by real values (reanalysis)

In this section, we study the effects of training the models using 'real' meteorological variables (*reanalysis*). The motivation is to check whether reducing the uncertainty of meteorological forecasts in the predictors help to obtain better models. However, results show that training with the 'real' meteorological variables seems to always appear the **overfitting problem**, especially with the method *Support Vector Machines*. It is tried to resolve this problem in *Random Forest* and *Gradient Boosting* by increasing the number of estimators (number of trees in the forest, *n_estimators*) and the minimum number of instances to grow a new leaf node (*min_sample_leaf*), but as the tables show, there are still quite some differences between the metrics of training set and testing set. It might be caused by the **dissimilarity in 2 data sets of forecasts and real values.** Since the results are not as good as trained with forecasts, we will not show further analysis but only the results (all the analysis and the results of *PCA* and *PLS* are on *Github*).

Unfortunately, using 'real' variables do not seem to help, contrary to our initial expectations.

4.2.1. Models trained by all variables

	LM	**RF (540 sec)**	**GB (3283 sec)**	**SVM (1974 sec)**
R^2 TR	91.5%	96%	99%	1
R^2 TE	**77.9%**	82.3%	**81.7%**	0
Max error TR	2770.8	1158.8	738.1	0.1
Max error TE	6925.8	1754.3	1984.7	2432.9
MAE TR	213	143.9	71.5	0.1
MAE TE	**4928.2**	276	**269.4**	802.9
RMSE TR	293.4	200.5	98.6	0.1
RMSE TE	**4947.1**	402.2	**399.8**	932.9

Table 4.7 The metric outputs using all the real time variables of Cadiz

4.2.2. Models trained by all variables except modules (new variables created)

	LM	**RF**	**GB**	**SVM**
R^2 TR	80.6%	94.9%	97.3%	98.6%
R^2 TE	**85.7%**	**83.5%**	81.4%	0

Max error TR	2756.2	1432.4	1032.4	1810.2
Max error TE	49295.5	1628.2	1723.2	2144.8
MAE TR	334.1	174.9	117.5	35.2
MAE TE	**44609.7**	**281.8**	287.9	929.3
RMSE TR	443.6	228.1	164.7	121.3
RMSE TE	**44627.3**	**382.8**	408.6	1022.4

Table 4.8 The metric outputs using all minus *modules* of real time variables of Cadiz

4.2.3. Models trained by all variables except u,v component variables

	LM	**RF**	**GB**	**SVM**
R^2 TR	89.6%	95.1%	97.9%	1
R^2 TE	**75%**	74.9%	**75.8%**	0
Max error TR	3520.9	1291.5	1074.3	0.1
Max error TE	7708.2	2062.9	2036.1	2432
MAE TR	234.7	158	102.8	0.1
MAE TE	**5653.8**	336.1	**318.7**	803.2
RMSE TR	325.6	222.5	144.9	0.1
RMSE TE	**5672.4**	476.8	**462.6**	933

Table 4.9 The metric outputs using all minus *u, v component* of real time variables of Cadiz

Chapter 5

Energy Forecasting for Spain

For the country, there are overall 442 combinations of coordinates, that is, there will be $442 \times 8 =$ **3536 predictors**. This amount of features makes the algorithm almost impossible to be run, therefore, dimension reduction application becomes compulsory.

On the following sections the ***PCA + GB*** and ***PLS + SVM*** metrics' outcome will be shown since the best results of dimension reduction for Cadiz are obtained from them. Furthermore, the relationship between features and response variables will be analyzed, as the following figure appears, it is reasonable to deduce that the most important features are derived from the coordinates where the most wind power installations are constructed.

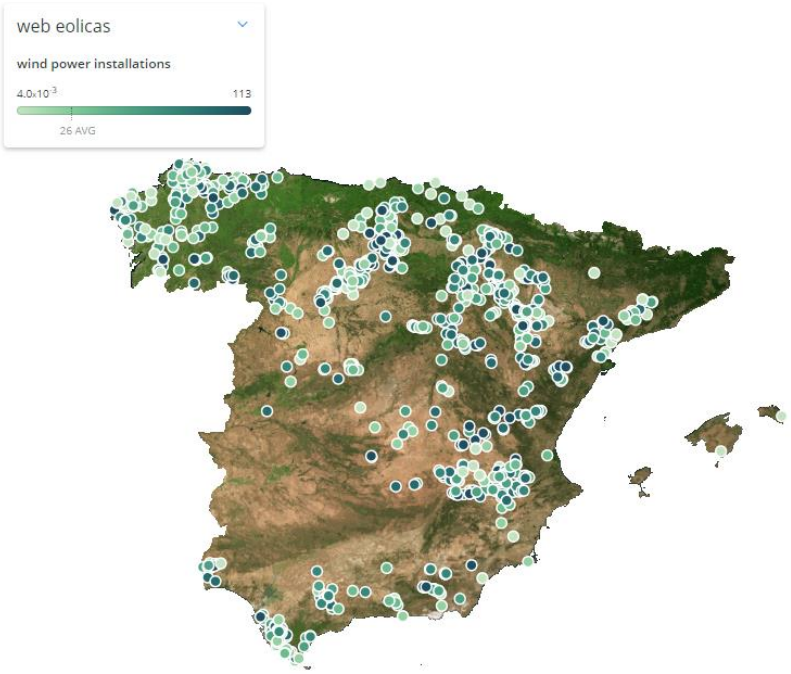

Figure 6.1 Map of national wind installations (Source: https://www.esios.ree.es/es/mapas-de-interes/mapa-instalaciones-eolicas)

5.1. Principal Component Analysis + Gradient Boosting

The very first step of training a model with *PCA* is to determine the number of components (new predictors that are composed by the original variables). The percentage of explained variance by the components is set to be a hyper-parameter for *PCA*, tuned together with the hyper-parameters of *Gradient Boosting.*

As demonstrated in the figure below, the machine has chosen 27 components. When the amount of the original variables increases, the fitting process becomes more complex. That causes the increment of the amount of the principal components so the decrease of the explained variance for each component.

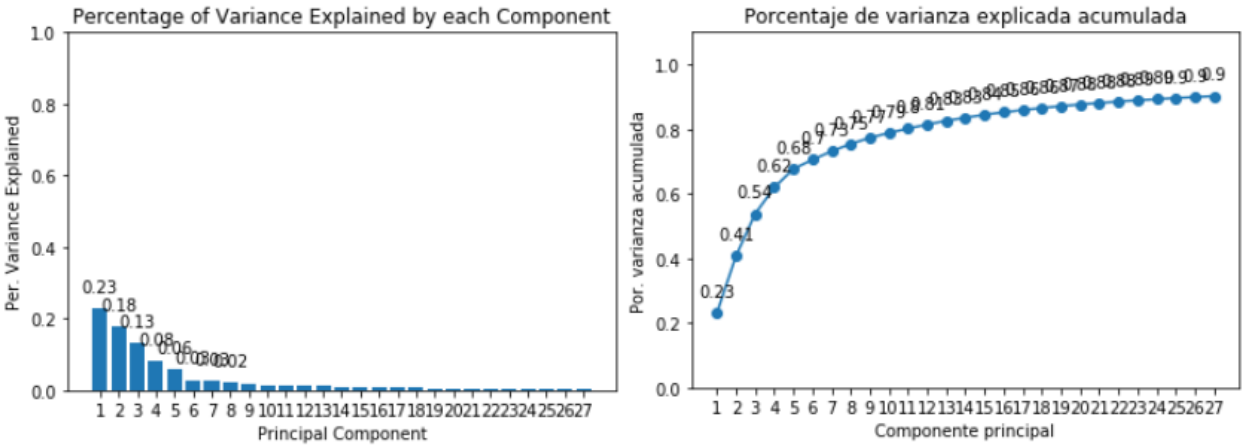

Figure 6.2 Output of Explained Variance of *PCA* of Spain

The following table presents the metrics' outcomes of *multiple linear regression* fitted with all the 3536 original variables in the first column; *multiple linear regression* fitted with the principal components (27) in the second column and the third column shows the metrics' values of *Gradient Boosting* trained with the 27 principal components.

Comparing the metrics, we can conclude that, in spite of that the time consumed *PCA + GB* is much more than *linear regression*, the model has not improved the estimation significantly. In addition, both *linear model* fitted will all variables and *PCA + GB* have overfitting problem as their *RMSE*s of testing sets differ remarkably from training sets although their explained variances are similar (setting *n_estimators* and *min_samples_leaf* larger or setting *max_depth* smaller had not solved the issue).

	LM_all_vars	PCA + LM	PCA + GB
R^2 TR	98%	90.6%	98%
R^2 TE	**92.5%**	91.3%	**92.5%**
Max error TR	184837	605391.6	176825.3
Max error TE	396570	368135.1	300926
MAE TR	21766.9	46641.4	21736.4
MAE TE	**44624**	47912.4	**43892**
RMSE TR	28097.5	61644.8	27927.7
RMSE TE	**57670.2**	61498.7	**57191.5**

Table 5.1 The metric output for *PCA* of Spain

5.2. Partial Least Squares + Support Vector Machines

Unlike *PCA*, it is unviable for *PLS* to tune the number of components as a hyper-parameter in a pipe. For this reason, we will fix the amount of components by resolving the optimization problem of maximizing the *Explained Variance* and minimizing the *Mean Square Error*.

Figure 6.3 shows the *R squareds* and *MSE*s of the corresponding regressions that are fitted with the number of components on the *x* axis. The *Red Cross* indicates the optimal point which is 39 in the case. Nevertheless, it is clear to notice that metrics' values had not changed much since *x* axis is around 20. In order to avoid overfitting, we will try modeling with 20, 30, and 39 components.

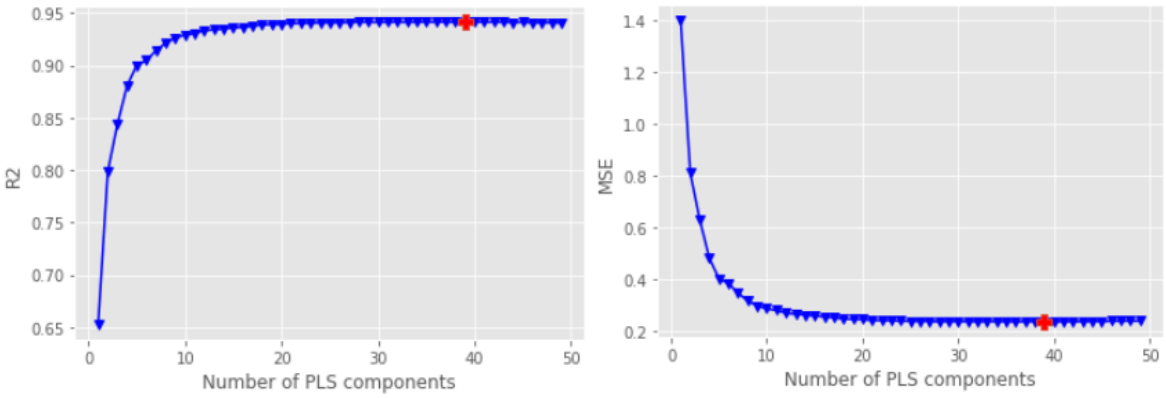

Figure 6.3 Output of Optimization for selecting the number of variables for Spain

The following table displays the metrics' results of *linear regression* with all variables; *PLS regresssion with 20 components and PLS + SVM* of 20, 30 and 39 components.

The best results are highlighted on the table, obtained with **combination of *PLS + SVM*** (20 components) with nearly **96% of explained variance** of testing set, and **mean absolute error of 31496.7**, comparing these metrics to other models' metrics, the results have become much better.

Furthermore, the differences between the metrics of training set and testing set are as hypothesized previously, the models with 30 and with 39 components have overfitted the training data. Mainly the *SVM* model with 39 components, the explained variance of the training set is almost 1, and the difference between *RMSE* is enormous. Models with more components increase the probability of overfitting.

	LM_all_vars	**PLS Regr.**	**PLS + SVM (20 comps)**	**PLS + SVM (30 comps)**	**PLS + SVM (39 comps)**
R^2 TR	98%	95%	**98%**	99.1%	99.9%
R^2 TE	**92.5%**	94.6%	**95.9%**	94.8%	93.4%
Max error TR	184837	476889.4	296944.1	201230	152171
Max error TE	396570	393823.9	371452.9	380258	445554
MAE TR	21766.9	32704.3	**24353.3**	10708.2	1798.9
MAE TE	**44624**	36731.4	**31495.7**	34383.5	37950.1
RMSE TR	28097.5	44208	**33193.8**	18507.2	6353.63
RMSE TE	**57670.2**	48674.2	**42435.7**	47644.2	53527

Table 5.2 The metric output for PLS of Spain

5.3. Removing Variables according to LM Coefficients

After having trained the models with *PCA + GB* and *PLS + SVM*, the target is to study and understand the relationship between the meteorological variables and the electricity production of Spain in order to prove the guess of that, the most important meteorological variables are from the coordinates (regions of Spain) that has most wind power installations.

With this intention, firstly, the *multiple linear regression* is fitted with all the variables. The following plot shows the coefficients of model. We can observe that, from all 3536 variables, most of them have coefficients close to 0, which probably implies that they would not make a significant difference on the final estimation of Spanish electricity production.

So the decision is to keep only the 500 out of 3536 variables that have the largest coefficients (in absolute value) in the *linear regression*, then fit them with the machine learning methods mentioned in chapter 3. Lastly, feature importance of the best model will be analyzed.

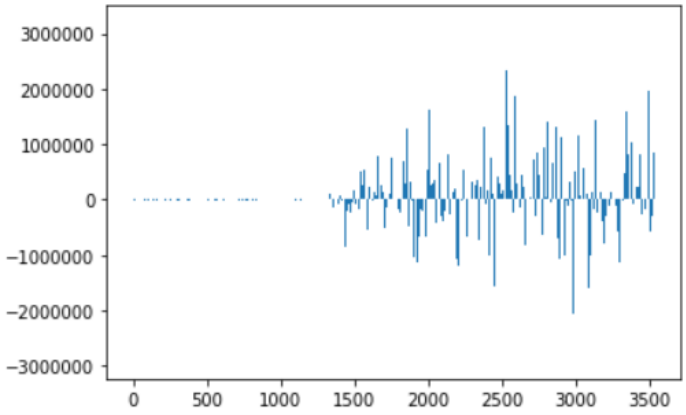

Figure 6.4 Plot of the coefficients of linear model fitted with all the variables for Spain

Gradient Boosting has fitted the best model according to the table below, despite that the model is a little worse than *PLS + SVM*, but it is significantly better than linear regression.

	RF	**GB**	**SVM**
$\boldsymbol{R^2}$ TR	84.4%	99.2%	99.9%
$\boldsymbol{R^2}$ TE	80.6%	**94.5%**	90.5%
Max error TR	385155	117210	105434,4
Max error TE	327288	415134	432082.4
MAE TR	63813.1	13962	490.2
MAE TE	74013.1	**35850.2**	46496.1
RMSE TR	79285.8	17899	3575.8
RMSE TE	91695.5	**48929.8**	64349.5

Table 5.3 The metric output fitted with 500 variables for Spain

- **Permutation Test**

As shown in the table below, the most weighted features according to *Gradient Boosting* fitted with 500 variables are, the ***wind speed moving towards the east, at 100 metres above the surface in coordinate (41,-1)***; ***the air speed at 100 metres towards the north of coordinate (42.5, -4)***. The

features that affect the final predictions the most are all related to the air speed at 100 metres above the Earth surface on different coordinates and directions.

Weight	*Feature*
0.0637 ± 0.0021	u100_41.0_-1.0
0.0323 ± 0.0010	v100_42.5_-4.0
0.0265 ± 0.0035	u100_42.0_-5.5
0.0181 ± 0.0018	u100_38.5_-1.5
0.0170 ± 0.0015	u100_37.5_-2.5
0.0164 ± 0.0021	v100_42.5_-1.0
0.0085 ± 0.0008	v100_41.0_-4.5
0.0067 ± 0.0010	v100_41.5_-1.5

Table 5.4 Output of Permutation Test of *Gradient Boosting* using 500 variables of Spainish meteorological variables

The following map displays that, the most important **coordinate (41.0, -1)** is in the city **Teruel**, the amount of aerogenerator installations around that coordinate is shown on the map, the dark colour implies that there are a lot of aerogenerators installed. Also, the **coordinate (42.5, -4)** is around **Burgos**, the city also contributes a lot of energy production with large amount of aerogenerator installations.

Figure 6.5 Map of the coordinate (41.0, -1)

- **Partial Dependence Plot**

PDP plot of 3 most important variables and 1 interactions according to the permutation test will be shown.

For the feature ***u100_41.0_-1.0,*** its increment causes more energy production, also, as the value of the **wind speed towards the east** becomes larger (around 67), the increment of the electricity production slows down.

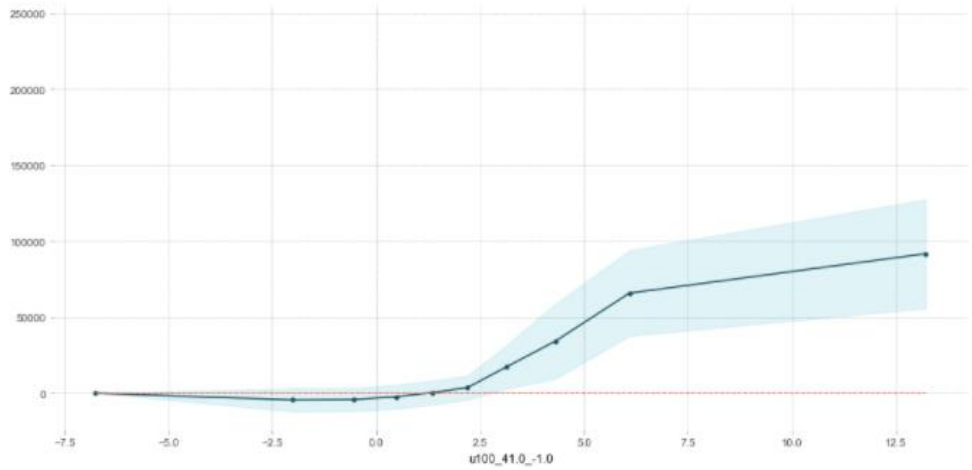

Figure 6.6 Output of Partial Dependence Plot of variable *u100_41.0_-1.0*

From the *PDP* plot of ***v100_42.5_-4.0***, as for Cadiz, we can observe that negative values of this variable favours the energy production, which implies that, the air moving toward the west favours the energy production.

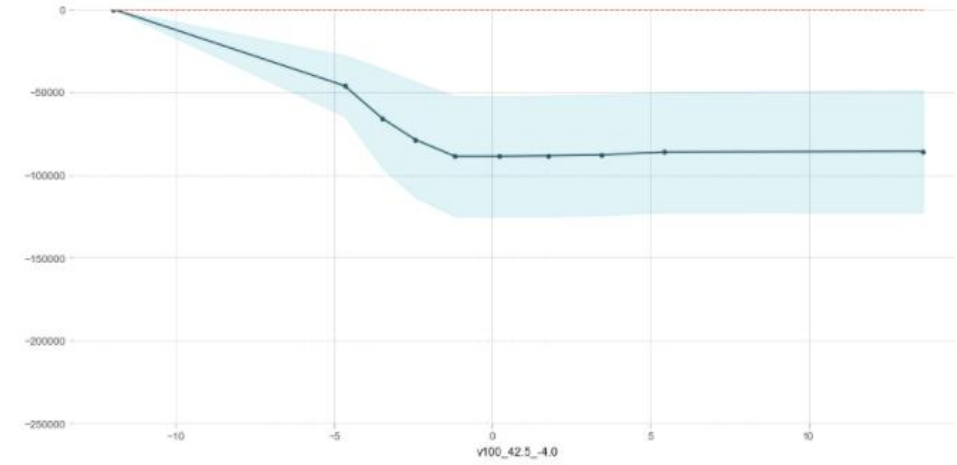

Figure 6.7 Output of Partial Dependence Plot of *v100_42.5_-4.0*

The following PDP plot for the feature ***u100_42.0_-5.5*** shows that, at the coordinate (42.0, -5.5), positive wind speed towards east causes the increment of energy production.

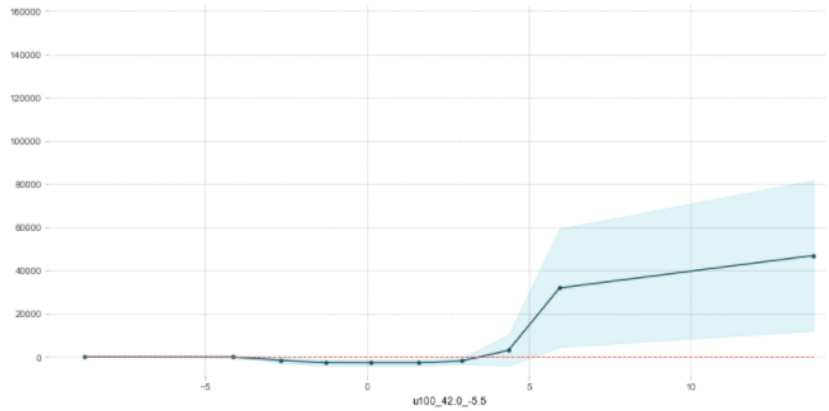

Figure 6.8 Output of Partial Dependence Plot of *u100_42.0_-5.5*

There is a pattern shown by all the single *Partial Denpendence* plots that, **higher wind speed toward east favours the energy production while higher wind speed towards the north disfavours**, which means that, the higher wind speed towards the west also favours the wind production.

The *PDP* interaction plot between variables *u100_41.0_-1.0* and *v100_42.5_-4.0* demonstrates that, the combination of higher *u100_41.0_-1.0* with lower *v100_42.5_-4.0* led a larger energy production (yellow area), and the other way around.

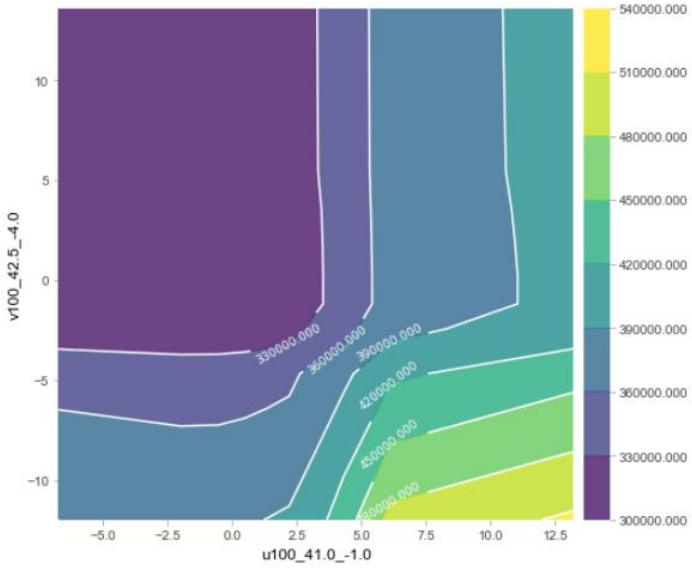

Figure 4.4 Output of Partial Dependence Plot of the interaction1

- **Summary Plots**

This plot shows that, larger value of *u100_41.0_-1.0* and *u100_42.5_-4* cause more wind energy production while larger value of *v100_42.5_-4.0* decreases the wind energy production. The corresponding warmer colour means that the variable causes the increment of the energy production with whether positive or negative values.

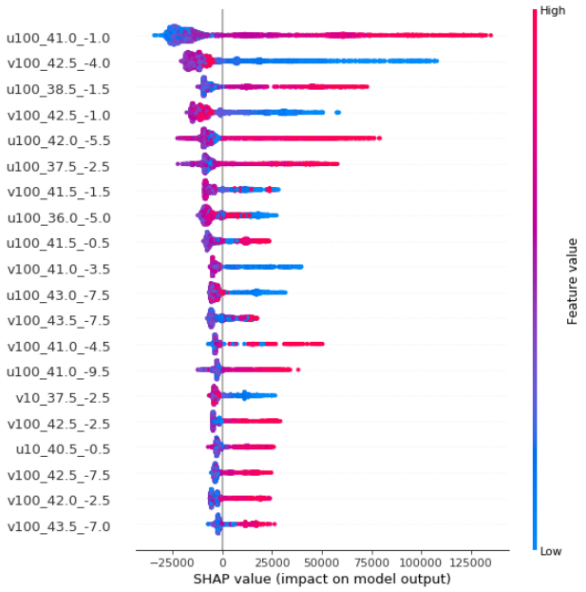

Figure 4.6 Output of Summary Plot for Spain

Chapter 6

Conclusion

Global warming is the long-term anthropogenically generated temperature increase of our planet's overall temperature. Due to the increase of electricity necessity of the human population, its pace has significantly quickened in the last hundred years as a result of the burning of fossil fuels.

Renewable energy has been an essential way of reducing greenhouse gas emissions, it efficiently increases electricity output worldwide. The rapid progress of renewable energy has led to an increasing proportion of such sources relative to standard resources; wind energy has been the most mature and developed renewable energy in Spain. Thus, the main aim of the project is to estimate the wind energy production of Cadiz and Spain applying machine learning tools using a 3-hour interval time series data set. The ML methods used are *Random forest, Gradient Boosting* and *Support Vector Machines* with their corresponding hyper-parameter tuning. In order to speed up the algorithm, dimension reduction techniques are also used. The predictors are the meteorological variables obtained in a spatial grid that covers the region of interest (Cadiz or Spain, in this work).

For the province **Cadiz**, despite the *Gradient Boosting* method has done a very nice job with mean absolute error of around 181.9 and explained variance of 90.9%, the **best model fitted is *Support Vector Machines* after applying *Partial Least Squares*** *transformation* with the forecasted data set, with a **mean absolute error of 179.9** and **explained variance of 91.2% with only 21 components, instead of 96.** The *PLS* has not only improved the error margings, but fastened the algorithms' speed significantly by reducing the dimension. According to the permutation test (that tests feature importance), the most important variables are mostly the new variables created (modules), which implies that **applying the technique of feature engineering has helped to improve the modeling process**. The most important coordinates for the province is **(36.5, -6.0),**

we can match it in the wind installations map seeing that there are quite some aerogenerators operating around that coordinate.

Moreover, compared to the forecasted predictors done by *ECMWF*, the **'real' meteorological variables tend to overfit the training data** in the case of Cadiz. As a consequence, we will only use the forecasted predictors for the estimation of Spanish electricity production.

For the entire country, only a few models are constructed due to the fact that the amount of predictors (3536 meteorological variables) makes the fitting process exceedingly computationally intensive for our available resources. In spite of that, the fitted models have performed reasonably well. Just as before, the best metrics are obtained from the combination ***PLS + SVM* with mean absolute error of 31495.7** and **explained variance of 95.9%,** it has improved outstandingly from the baseline model with only **27 components instead of 3536 original variables,** which is 0.7% (multiple linear regression fitted with all the 3536 variables has obtained a MAE of 44624 and R squared of 92.5%).

It should be noticed that the large difference between the error measures for Spanish electricity estimation and Cadiz electricity estimation are caused by the scale of their response variables. The Spanish electricity production is way higher than Cadiz electricity production. The explained variances for both regions are very similar.

Another interesting fact of modeling is that, as the method used for hyper-parameter space searching is random search implemented in *Python*, this function only chooses 10 random combinations of hyper-parameters. This is why as the complexity of the data set (dimension) increases, the precision of the results will decrease as the methods need to make more hyper-parameter combinations. For example, we can observe that *Support Vector Machines* always had the problem of overfitting, but it performed remarkably well after applying *Partial Least Squares transformation*, as well for *Gradient Boosting*. Namely, when fitting the models, we are also seeking that the machine will luckily randomly choose one of the most suitable union of hyper-parameters for the corresponding models. Of course, Random Search might improve its performance if given more computing resources than the ones available for this work.

Finally, there is still a possibility of improvement while estimating the renewable energy production, the task would probably remain a very computationally expensive process.

Biliography

Adrian Stetco, F. D. (2019, April). *Machine learning methods for wind turbine condition monitoring: A review*. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S096014811831231X#bib1

Ahmed, B. (2013, July). *An Analytical Study for Establishment of Wind Farms in Palestine to Reach the Optimum Electrical Energy.* Retrieved from Research Gate: https://www.researchgate.net/publication/334729294_An_Analytical_Study_for_Establishment_of_Wind_Farms_in_Palestine_to_Reach_the_Optimum_Electrical_Energy

Asociación Empresarial Eólica. (n.d.). *La eólica y sus ventajas*. Retrieved from Asociación Empresarial Eólica: https://www.aeeolica.org/sobre-la-eolica/la-eolica-y-sus-ventajas

Becker, D. (n.d.). *Machine Learning Explainability*. Retrieved from Kaggle: https://www.kaggle.com/learn/machine-learning-explainability

Bergstra, J. S. (2011). *Algorithms for Hyper-Parameter Optimization.* Retrieved from Google Scholar: https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

ECMWF. (n.d.). *ECMWF*. Retrieved from Copernicus: https://cds.climate.copernicus.eu/#!/home

Energías Renovables. (2021, March 12). *Las renovables han producido en España en 2020 casi el doble de electricidad que la nuclear*. Retrieved from Energías Renovables: https://www.energias-renovables.com/panorama/espana-ha-producido-en-2020-mas-electricidad-20210312

Ethem, A. (2020). *Introduction to Machine Learning.* Mit Press.

Gobierno de España. (2009). *Energías renovables*. Retrieved from Ministerio para la Transición Ecológica y el Reto Demográfico: https://energia.gob.es/desarrollo/EnergiaRenovable/Paginas/Renovables.aspx

James, G. W. (2013). *An introduction to statistical learning.* New York: Springer. Retrieved from Eoliccat: https://eoliccat.net/

Kuhn, M. &. (2013). *Applied predictive modeling.* New York: Springer.

Python - Scikit Learn. (n.d.). *Importance of Feature Scaling*. Retrieved from Scikit Learn: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html#sphx-glr-auto-examples-preprocessing-plot-scaling-importance-py

Python - Scikit Learn. (n.d.). *Pipelining: chaining a PCA and a logistic regression.* Retrieved from Scikit Learn: https://scikit-learn.org/stable/auto_examples/compose/plot_digits_pipe.html

Raman Arora, A. C. (2012). *Stochastic optimization for PCA and PLS.* Chicago: IEEE.

Red Eléctrica de España. (2021, March 12). *2020, el año de la energía más 'verde' gracias al récord en generación eólica y solar fotovoltaica*. Retrieved from Red Eléctrica de España: https://www.ree.es/es/sala-de-prensa/actualidad/nota-de-prensa/2021/03/2020-energia-mas-verde-gracias-record-eolica-y-solar-fotovoltaica

Red Eléctrica de España. (n.d.). *Sistema de Información del Operador del Sistema*. Retrieved from Sistema de Información del Operador del Sistema: https://www.esios.ree.es/es

Rodrigo, J. A. (2020, December). *PCA con Python.* Retrieved from cienciadedatos.net: https://www.cienciadedatos.net/documentos/py19-pca-python.html

Vung, P. V. (2019). *Partial Least Squares Regression in Python*. Retrieved from Kaggle: https://www.kaggle.com/phamvanvung/partial-least-squares-regression-in-python

Appendix

Github Repository: https://github.com/danyuz/TFM_energy.git