Machine Learning project
Global Data on Sustainable Energy dataset

Doan-Kien THAI, Jonas MELANDSOR, Elsa DANH-NGHET
4MA, 2024

INSA TOULOUSE

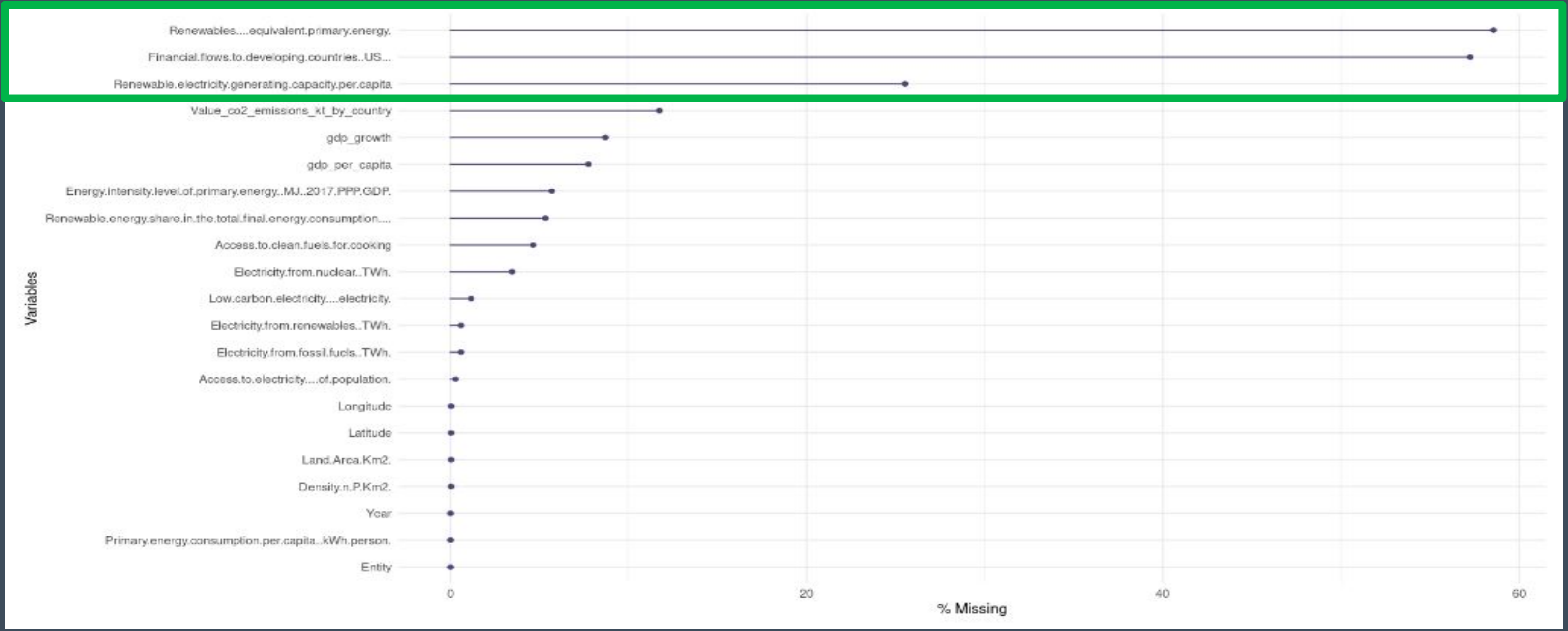
Table of contents

01

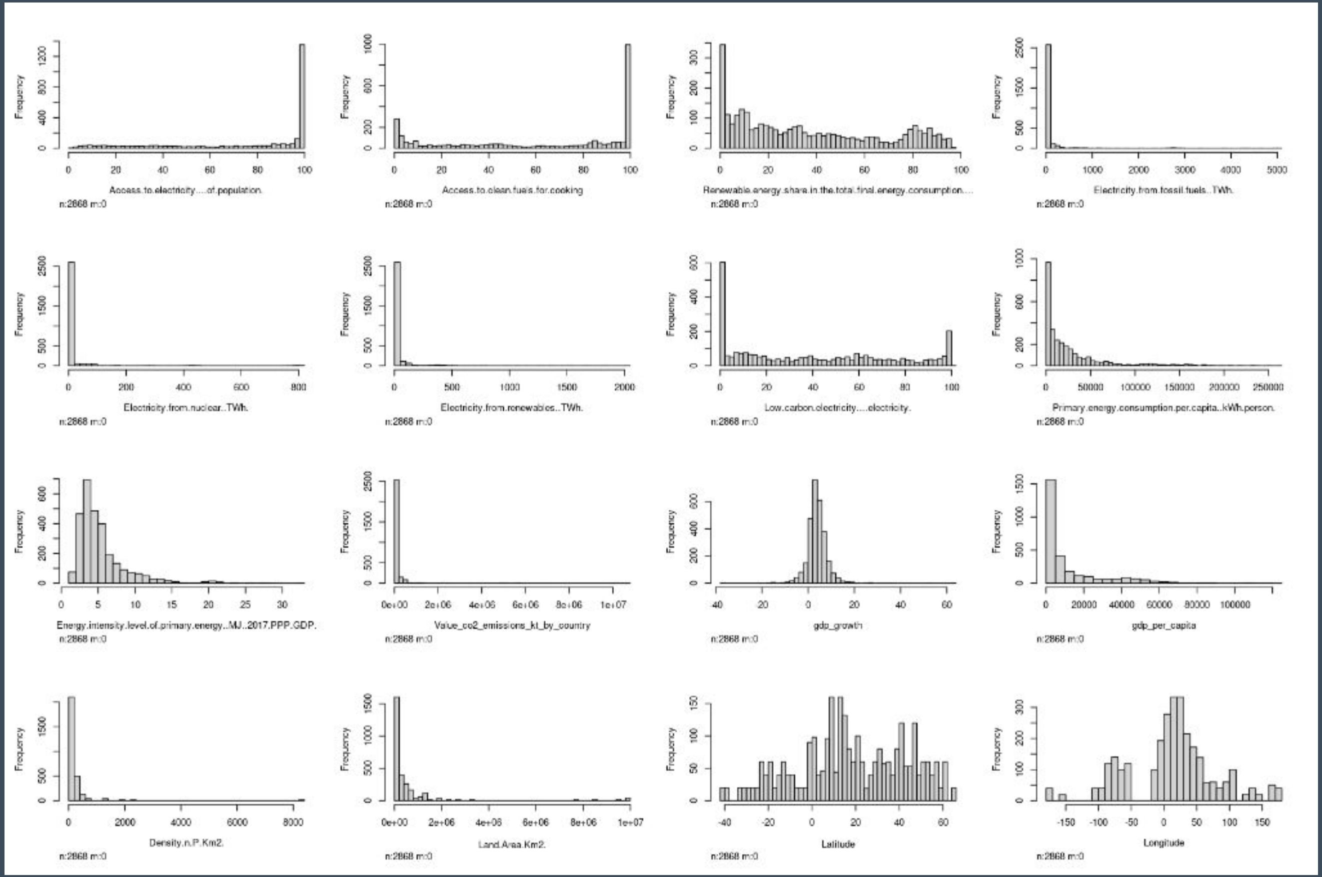
Exploratory Data Analysis

The dataset

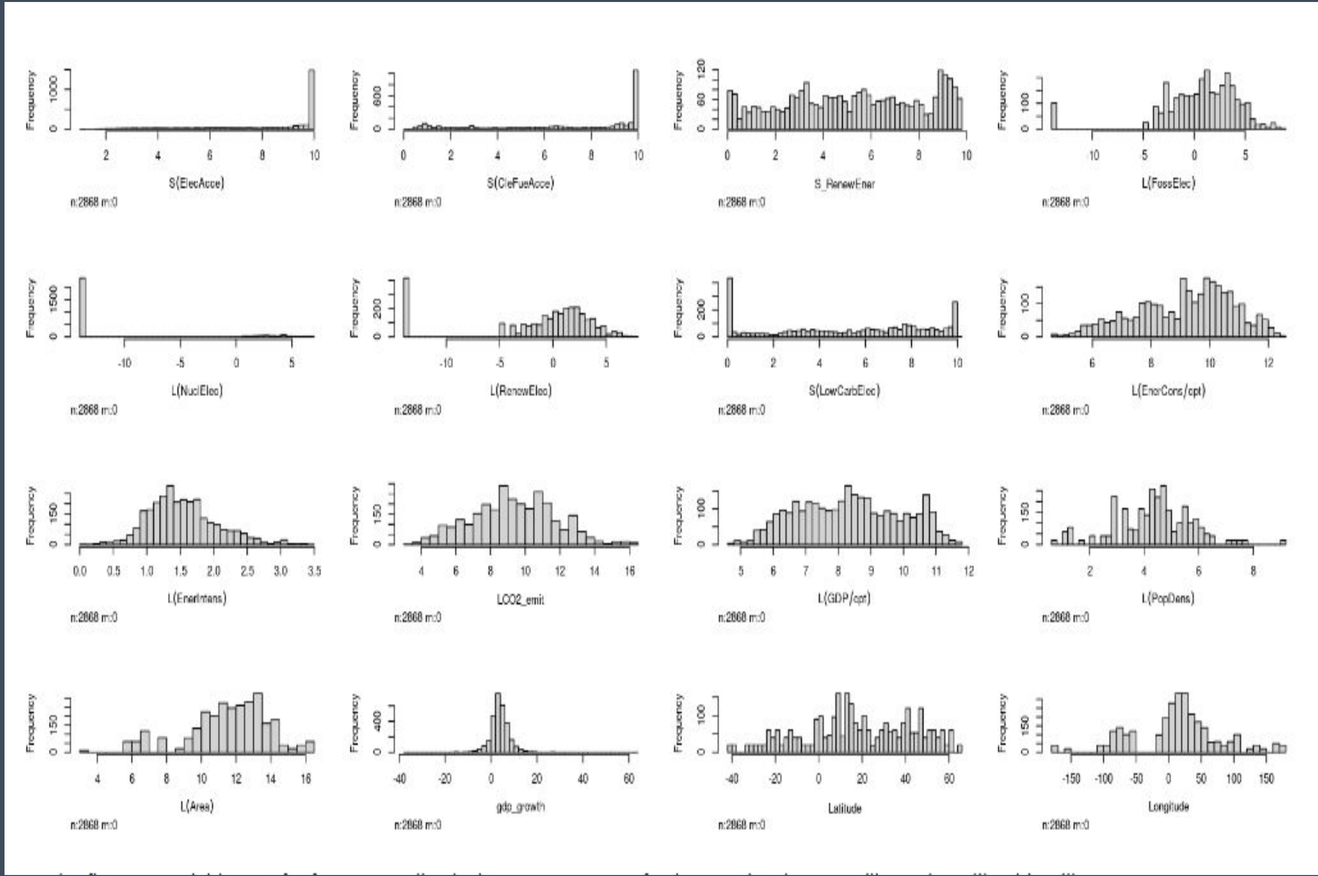
- 3649x21
- Removing variables and missing values



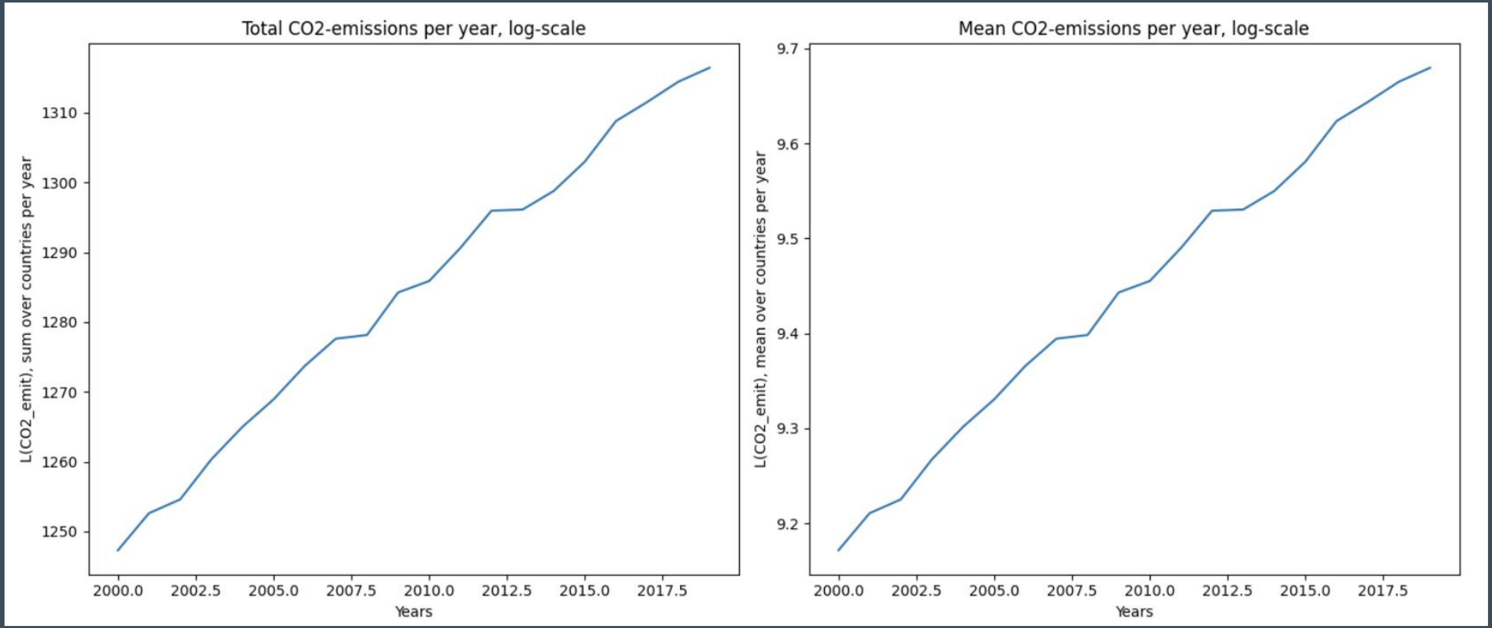
Unidimensional descriptive (1)



Unidimensional descriptive (2)

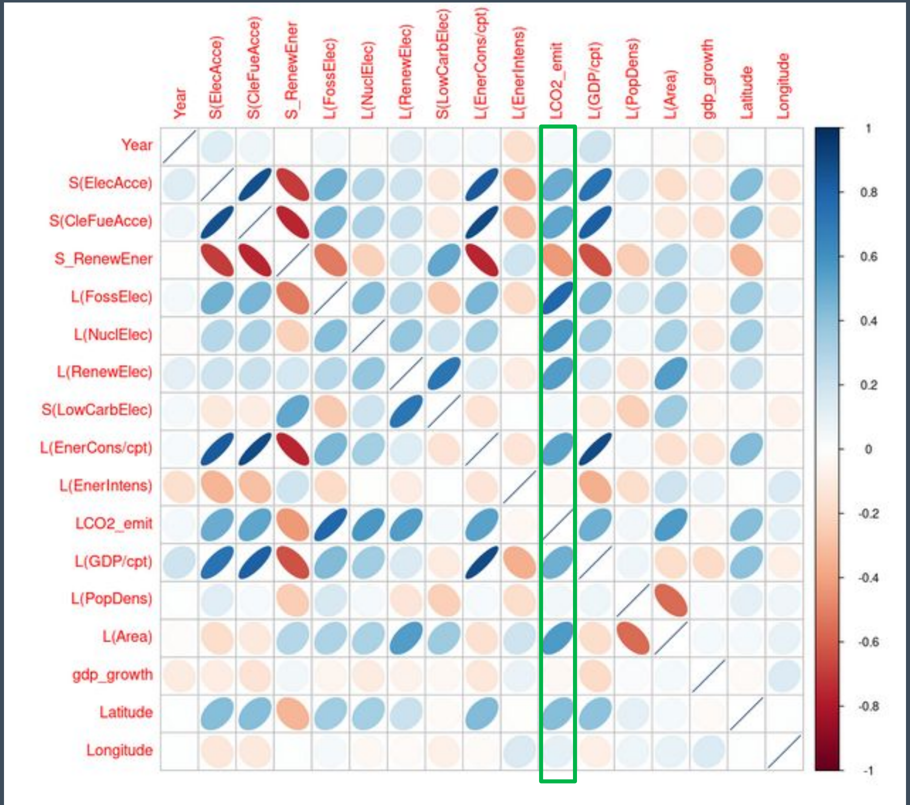


Heterogeneity and year

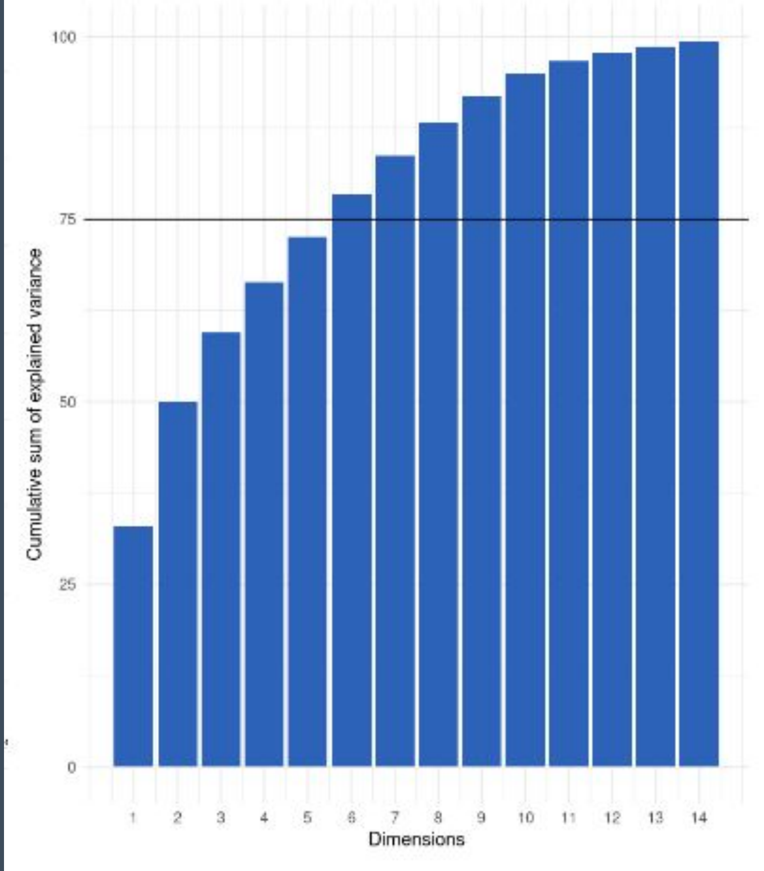


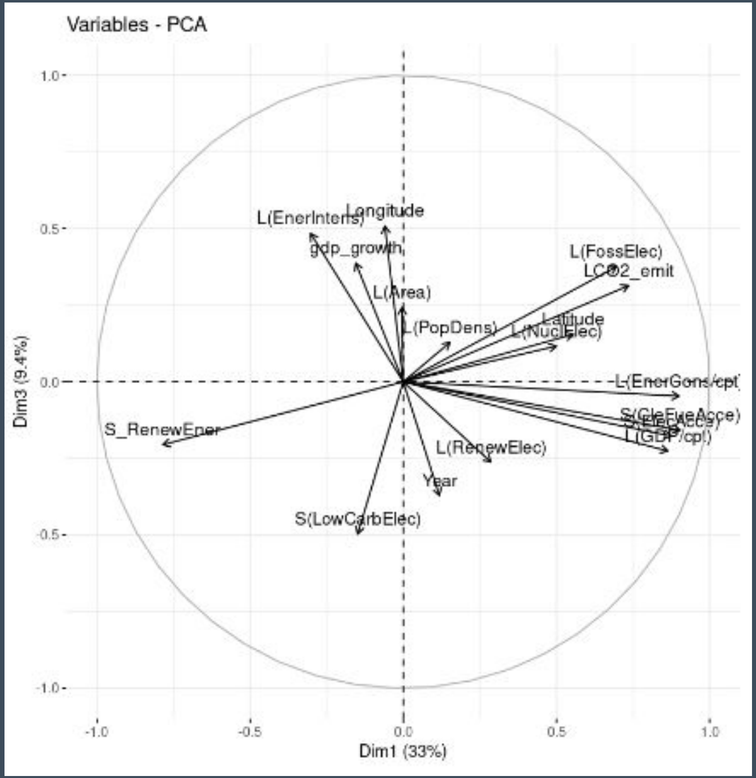
Correlation

- Electricity from fossil fuels → high correlation
- High correlation between some variables (Access to clean fuels for cooking and to electricity)
- Some variables seem to have low correlation with every variables



Principal Component Analysis

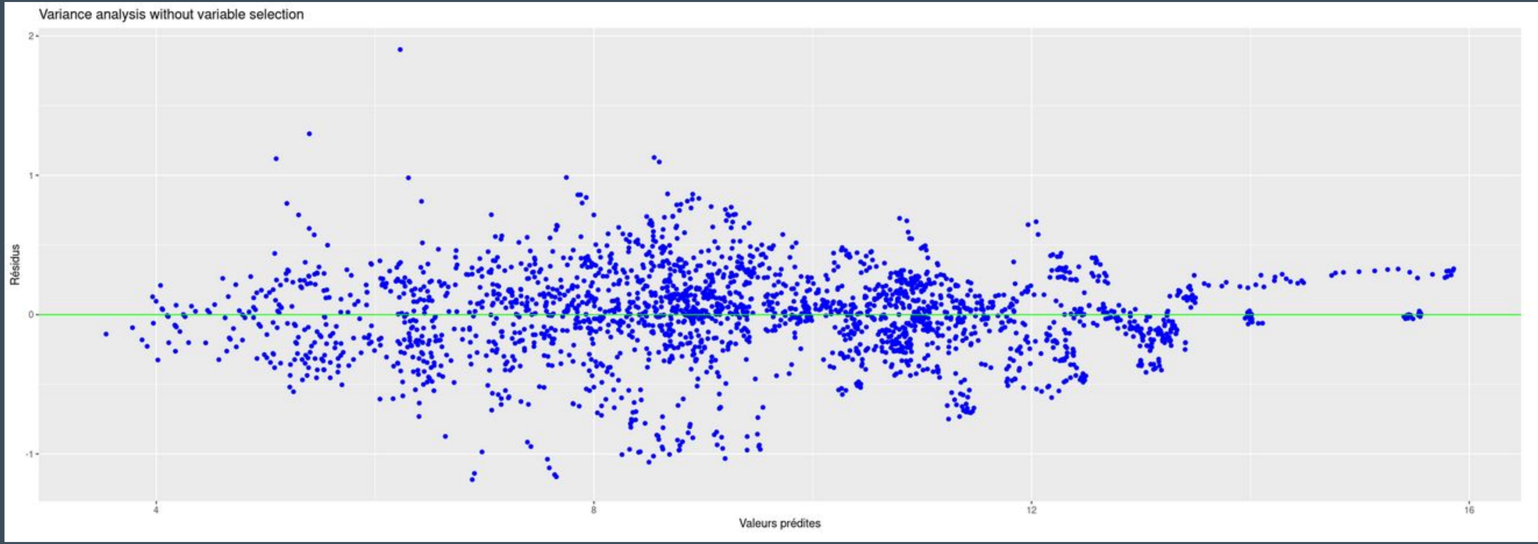




02

Methods Of Modelisation

Linear regression



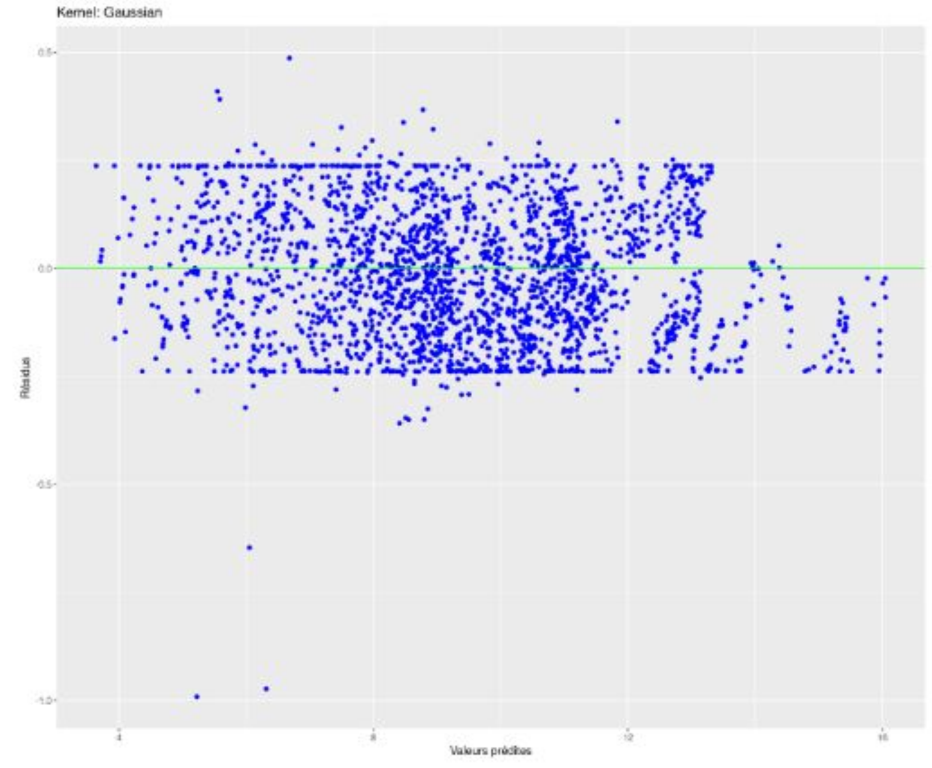
MSE :

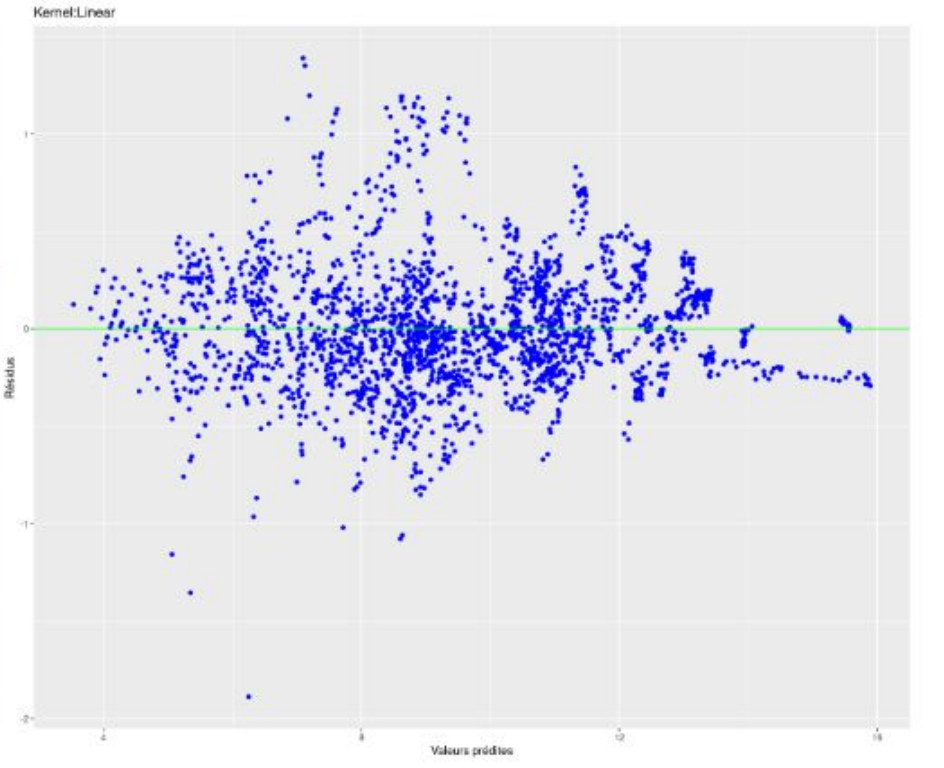
- Linear model without selection : 0.0998
- LASSO with λmin : 0.1006

SVR – Choice of kernel

Kernel	# Support Vectors	Parameters to tune
Radial/Gaussian	429	cost, gamma
Linear	869	cost
Polynomial	944 (degree 3)	cost, gamma, coef0, degree
Sigmoidal	2287	cost, gamma, coef0

SVR - Results (1)





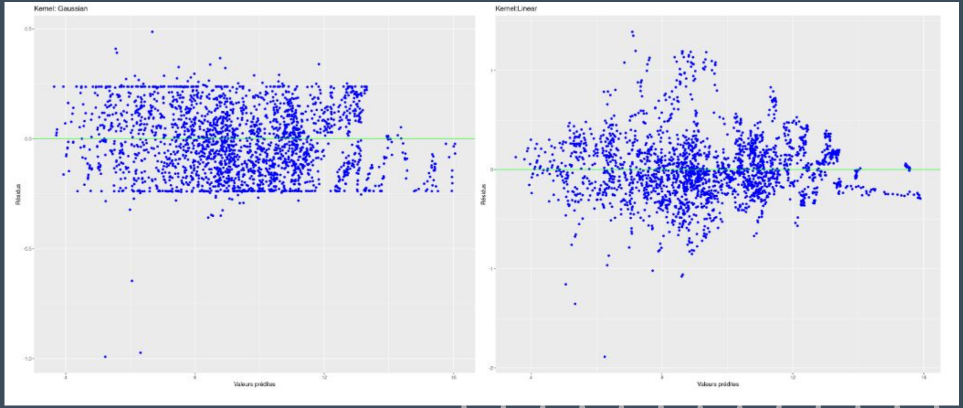
SVR - Results (2)

Radial kernel:

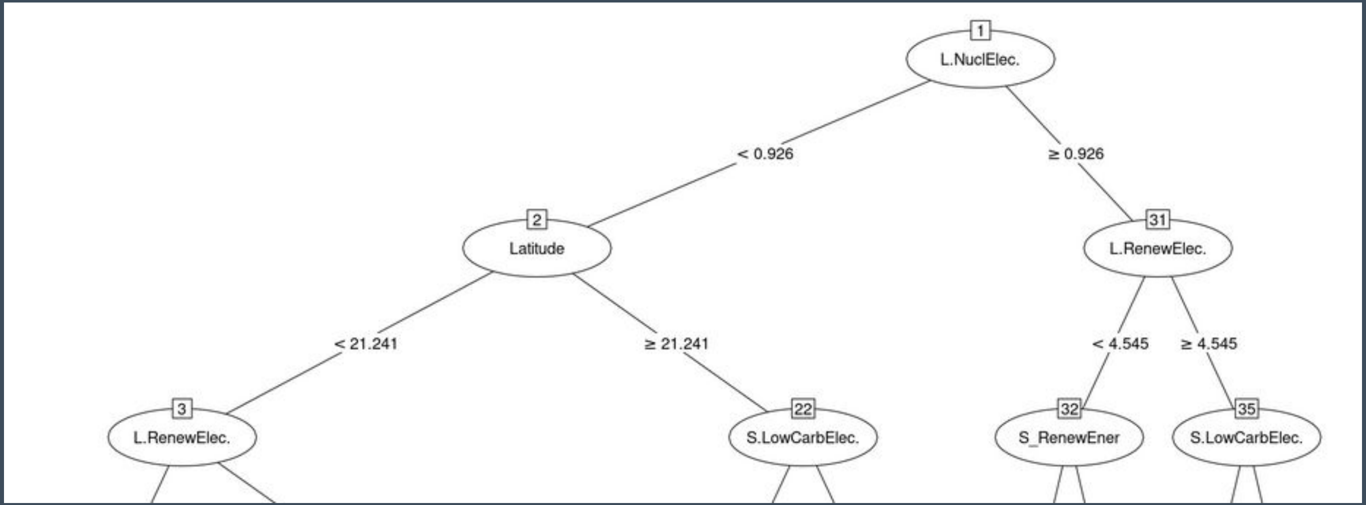
- Cost: 13
- # Support vectors: 290
- Error on training set: 0.0230
- MSE on test test: 0.0511
- Adjusted R^2: 0.9997

Linear kernel:

- Cost: 9
- # Support vectors: 873
- Error on training set: 0.1017
- MSE on test test: 0.1044
- Adjusted R^2: 0.9995



Classification and Regression Trees (1)



MSE :

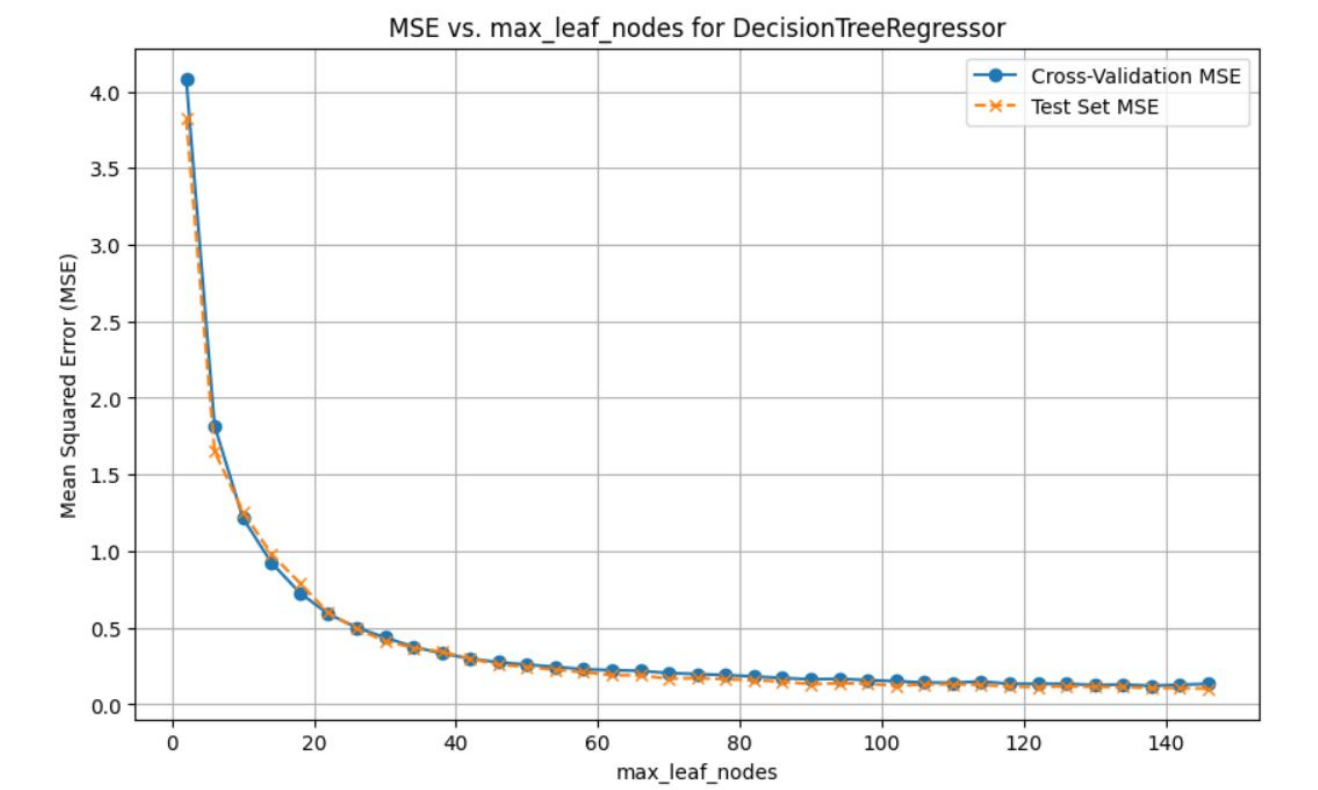
- ➢ Library rpart : 0.2693
- ➢ Library caret : 0.7807

Residuals by leaves

Variables of importance :

- ➢ Electricity from nuclear : high correlation with Value CO2 emissions
- ➢ Low carbon electricity : low correlation

Classification and Regression Trees (2)



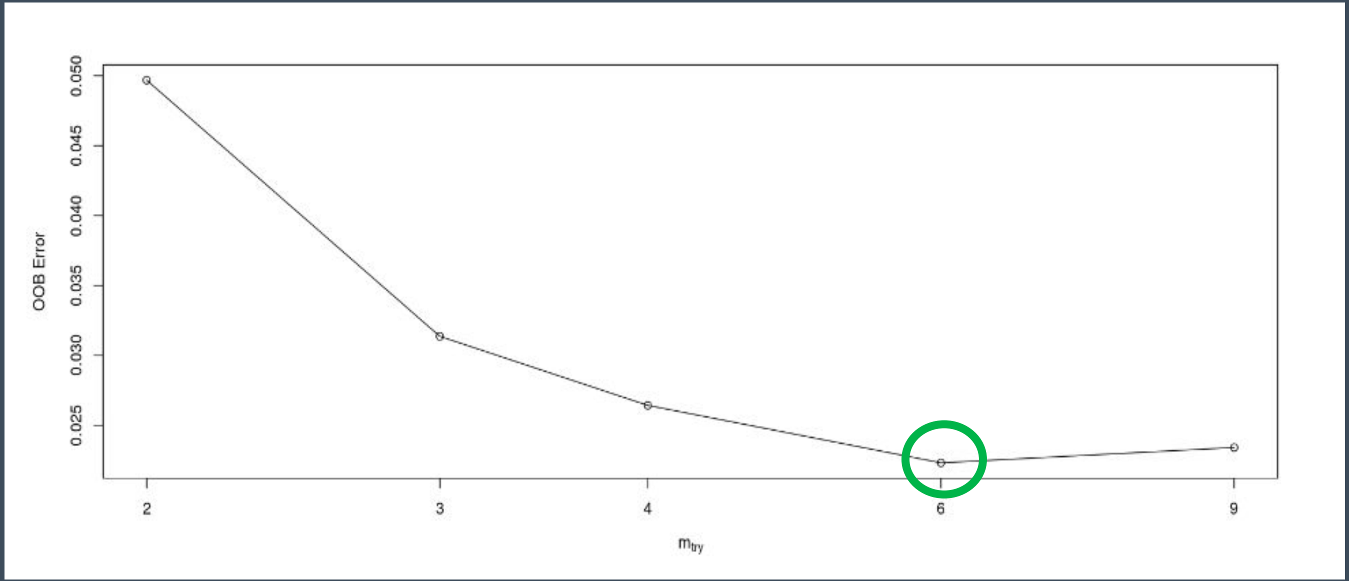
Random Forest - tuning (1)

- Tune mtry
- Choose number of trees

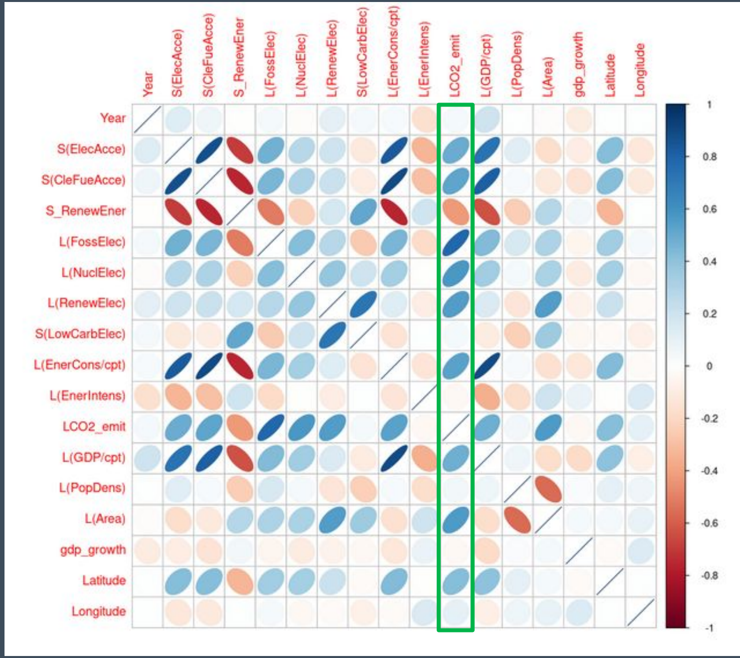
Tree	Out-of-bag MSE	Out-of-bag %Var(y)	Test set MSE	Test set %Var(y)
50	0.0236	0.42	0.01884	0.32
100	0.0196	0.35	0.01796	0.31
150	0.01864	0.33	0.01797	0.31
200	0.01845	0.33	0.0178	0.31
250	0.01799	0.32	0.01784	0.31
300	0.01786	0.32	0.0181	0.31
350	0.01788	0.32	0.01841	0.32
400	0.01762	0.31	0.01801	0.31
450	0.01764	0.31	0.01777	0.31
500	0.01743	0.31	0.01775	0.31

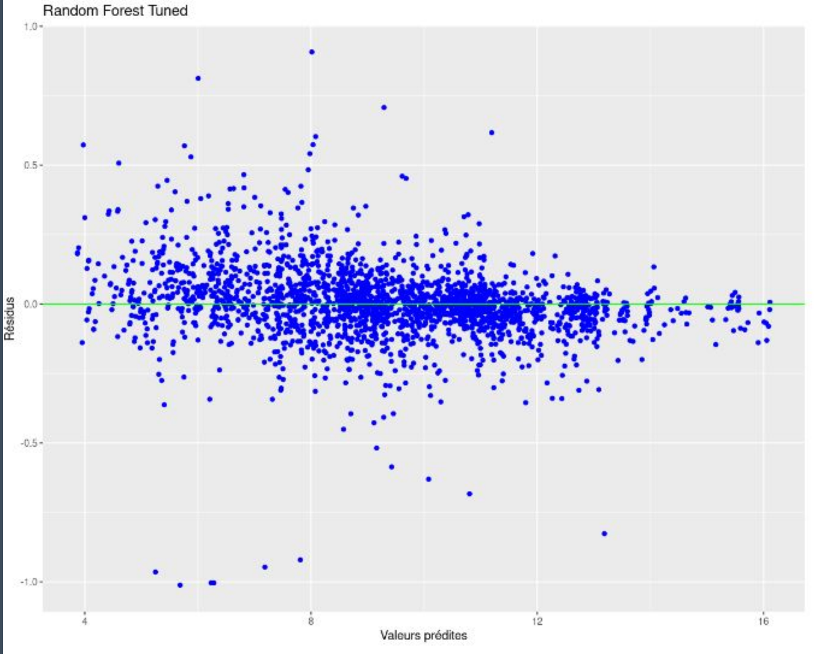
Random Forest - tuning (2)

- Tune mtry
- Choose number of trees

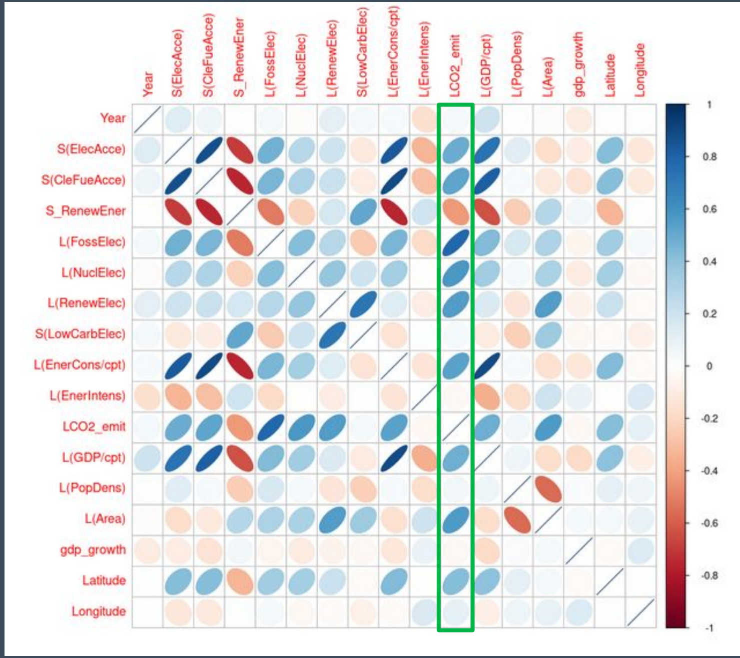


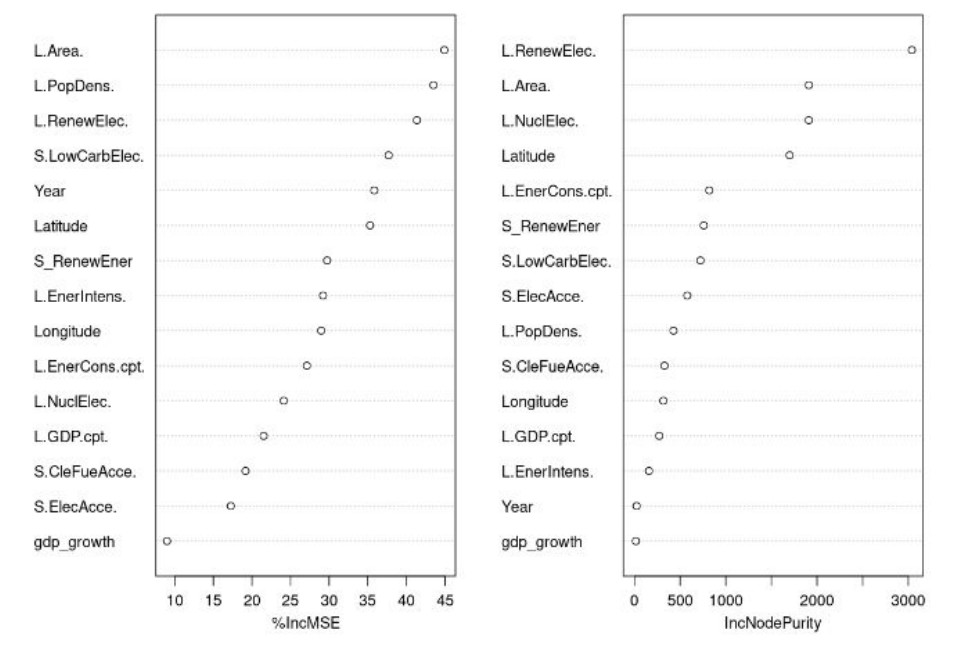
Random Forest - results (1)





Random Forest - results (2)

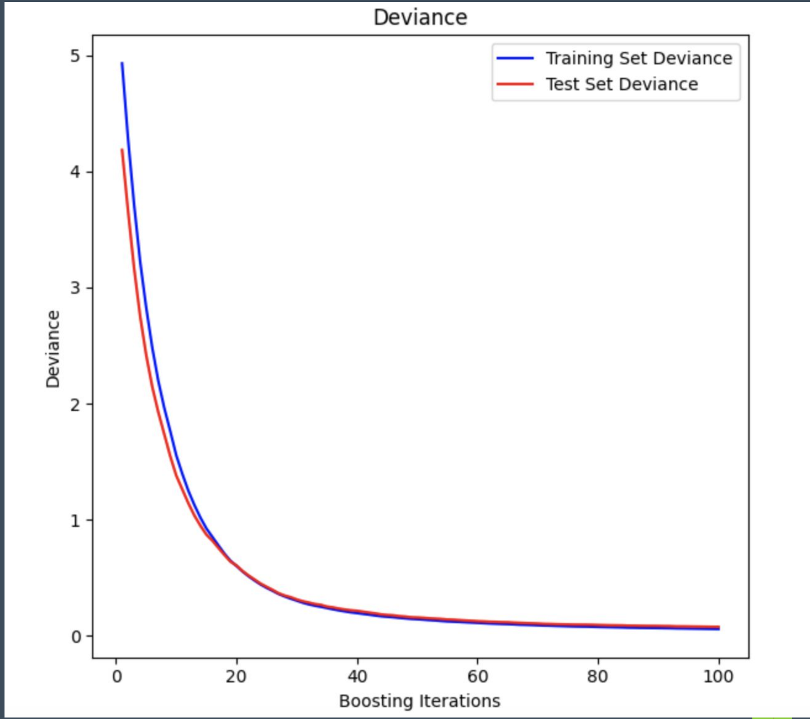




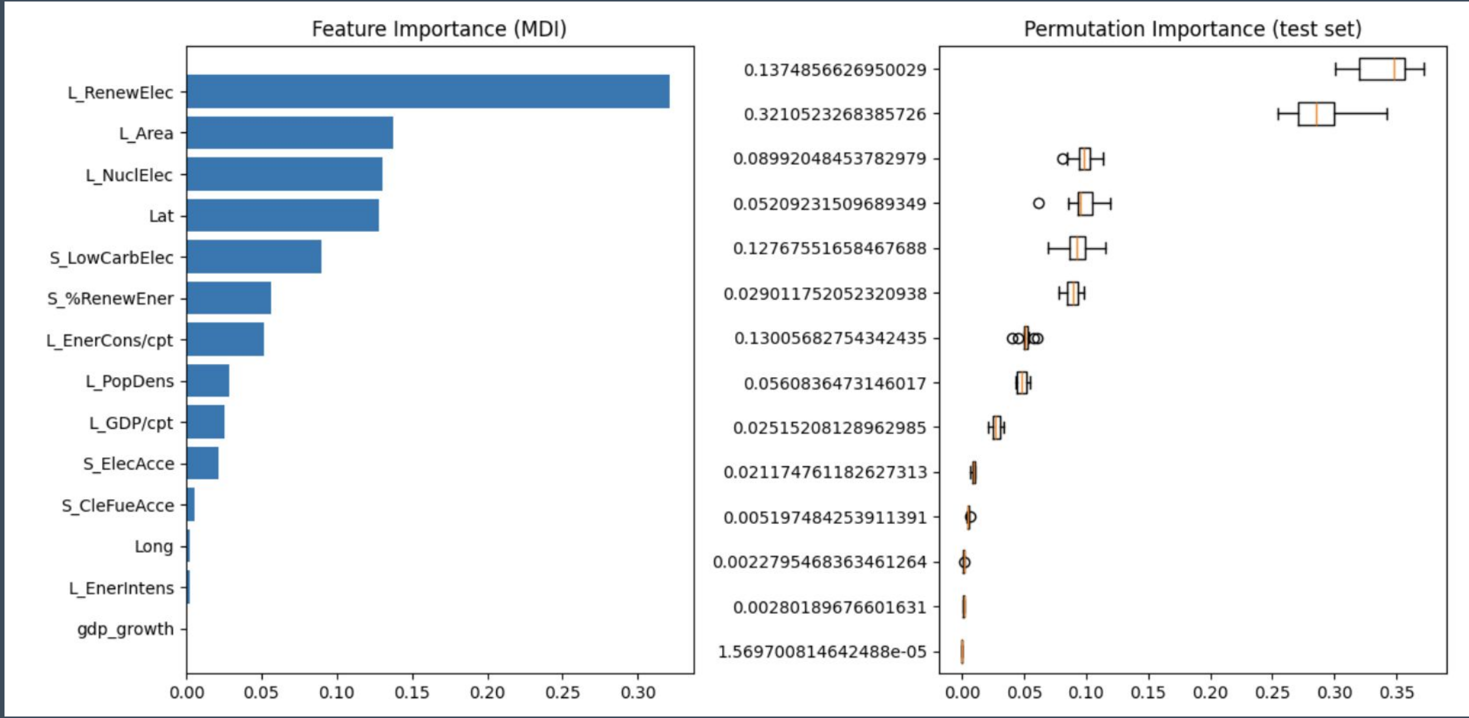
Boosting

```python
# Default parameter
params = {
    "n_estimators": 100,
    "max_depth": 3,
    "min_samples_split": 2,
    "learning_rate": 0.1,
    "loss": "squared_error",
}
```

```
The mean squared error (MSE) on test set: 0.0789
```



Boosting's feature importance



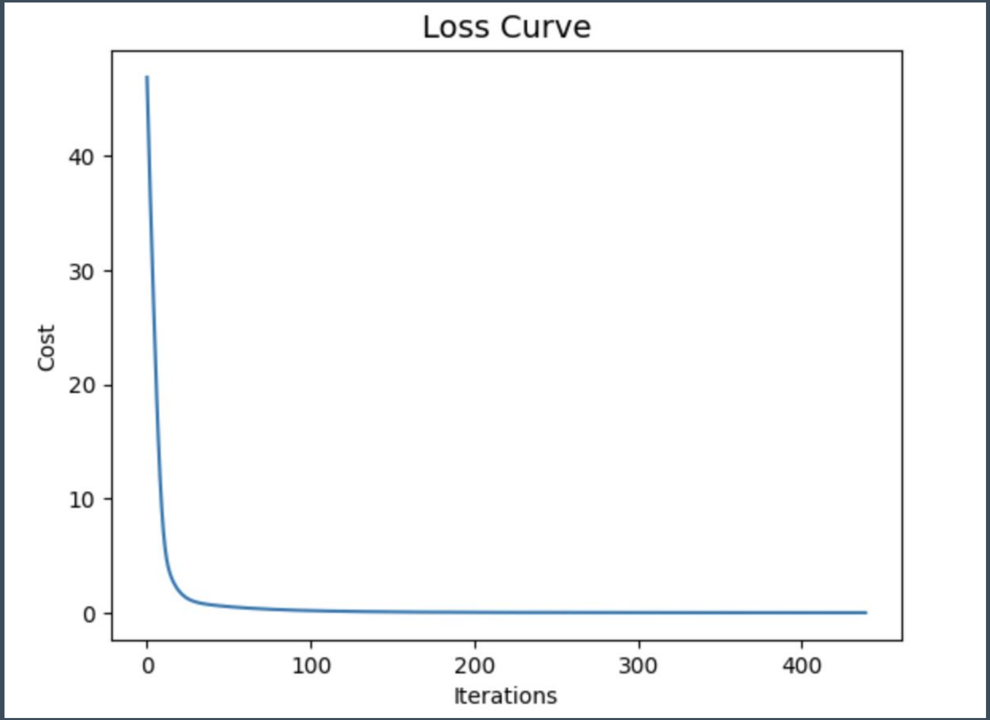
Boosting's tuning with GridSearchCV

```
grid['n_estimators'] = [10, 50, 100, 500]
grid['learning_rate'] = [0.001, 0.01, 0.1, 1.0]
grid['subsample'] = [0.5, 0.7, 1.0]
grid['max_depth'] = [3, 7, 9]
```

```
Best: 0.995694 using {'learning_rate': 0.01, 'max_depth': 9, 'n_estimators': 500, 'subsample': 0.5}
```

```
MSE:   0.019249866681743333
```

Neural networks



Neural Network's tuning

```
param_grid={"hidden_layer_sizes":list([(5,),(6,),(7,),(8,)])}
```

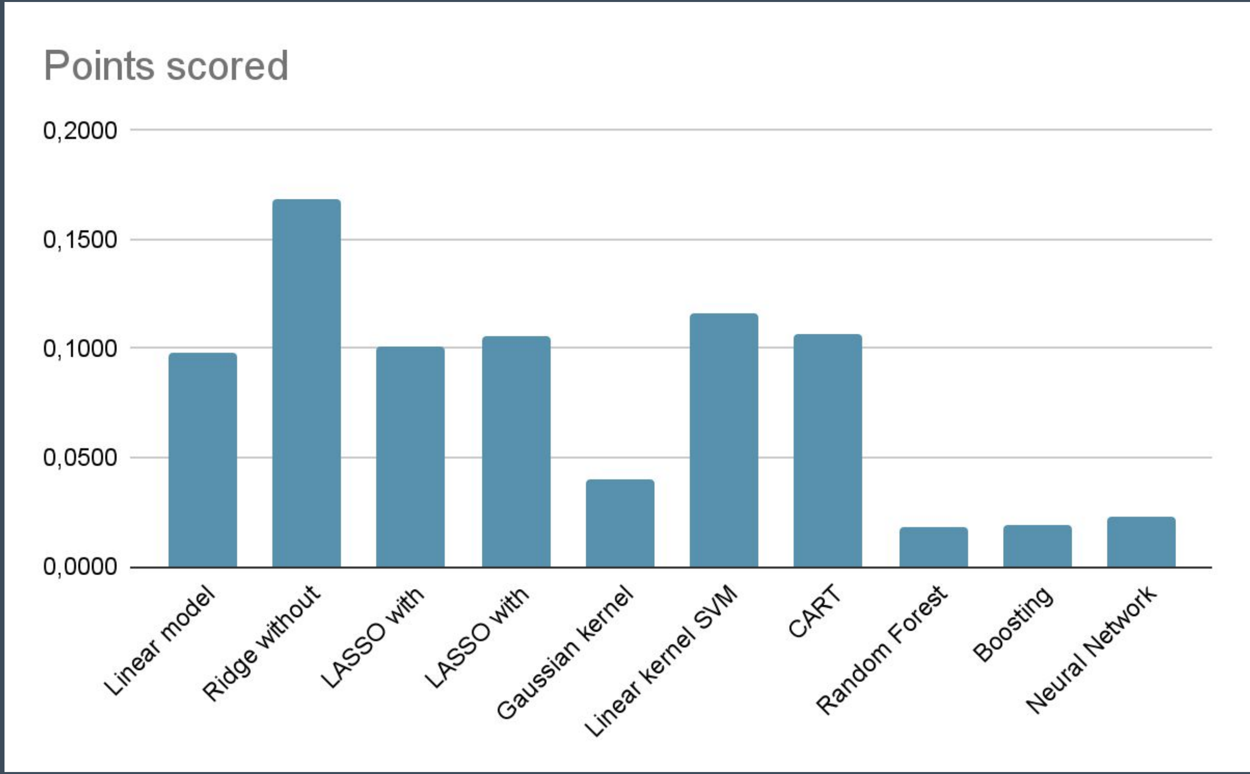
```
Best: 0.867450 using {'hidden_layer_sizes': (7,)}
0.796815 (0.053024) with: {'hidden_layer_sizes': (5,)}
0.704994 (0.353057) with: {'hidden_layer_sizes': (6,)}
0.867450 (0.041294) with: {'hidden_layer_sizes': (7,)}
0.812792 (0.041786) with: {'hidden_layer_sizes': (8,)}
```

```
MSE:   0.02398911369809242
```

03

Results And Comparison

Errors on test sample

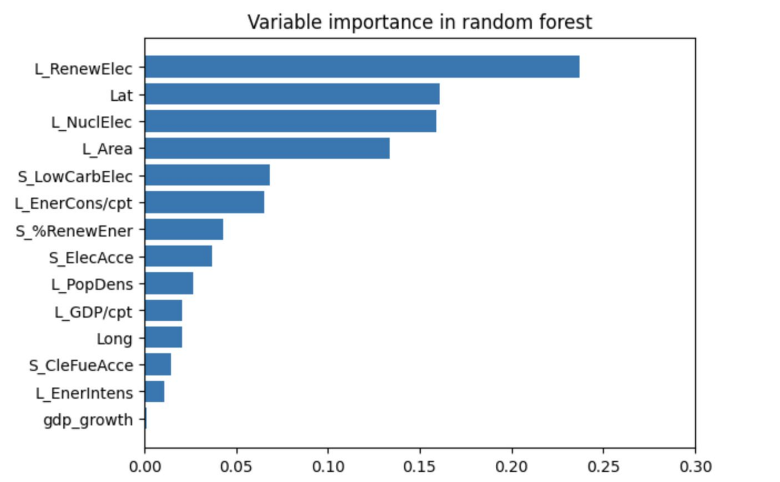


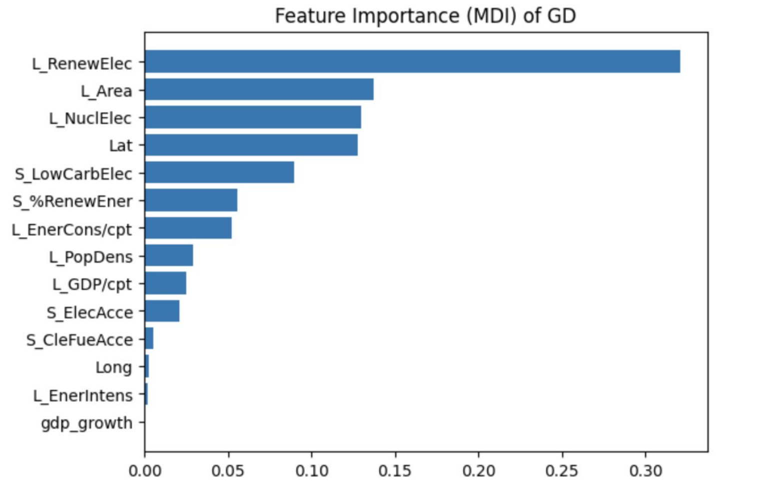
computation

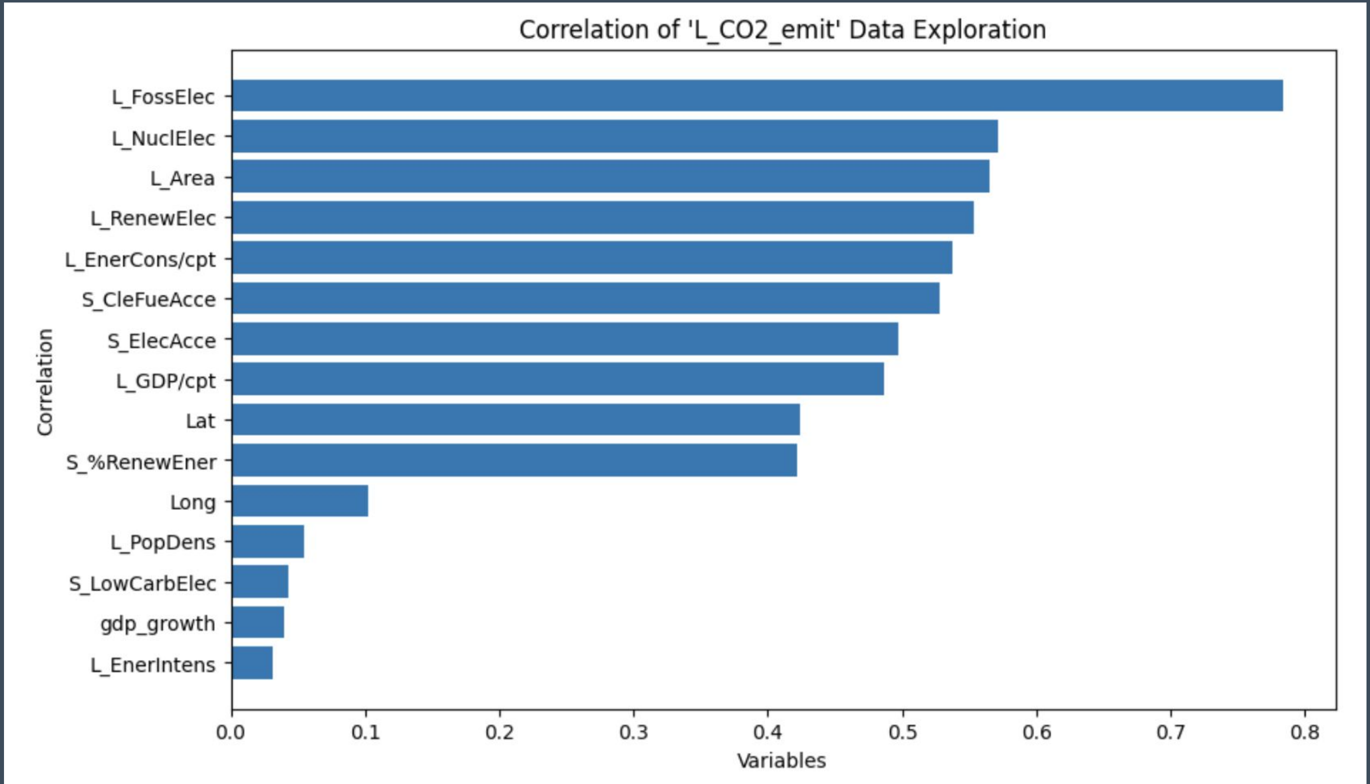
Methods	Time (s)
Linear (naïve)	54.4
Linear (Ridge)	111.1
Linear (Lasso min)	112.6
Linear (variable selection)	54.8

SVR (radial)	1185.111
SVR (linear)	1618.473
CART	1633.389
Random Forest	1675.495
Boosting	1755.166
Neural network	8406.133

Comparison modelisation and data analysis expectation



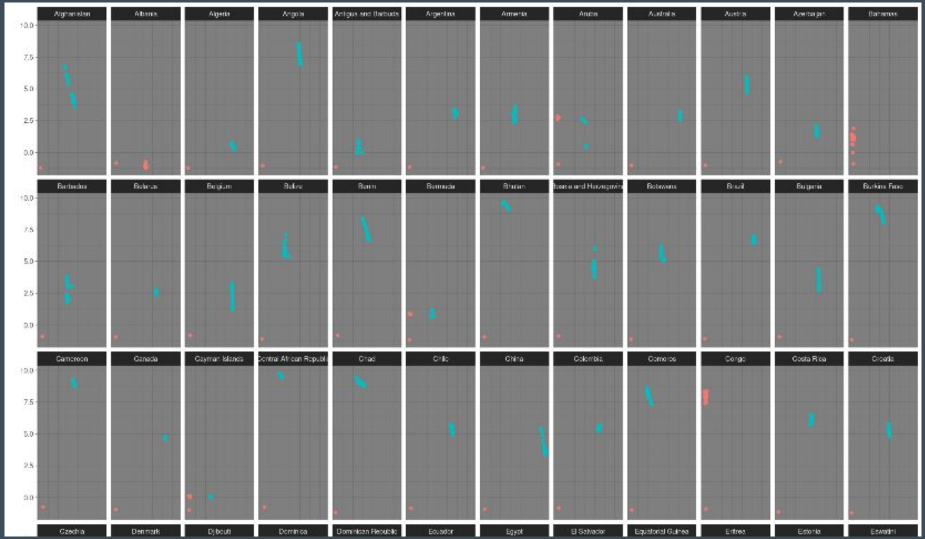




We can see there are some differences in variables' importance. The Random Forest Regressor has valued the variance "S_LowCarbElec" significantly, while assessed the variance of GDP growth as almost uninfluenced.

Missing values - Method

- Transforming data
- Not removal of L(FossElec)
- MAR or MNAR



Methods of imputation:

- LOCF
- By mean
- By median
- KNN
- MissForest
- Amelia II

Missing values - Results

Linear regression	
Imputation methods	MSE on test sample
LOCF	0.3242
Mean	0.7427
Median	0.7283
KNN	0.4154
MissForest	0.1142
Amelia	0.0847

Random Forest	
Imputation methods	MSE on test sample
LOCF	4.034
Mean	4.067
Median	4.750
KNN	4.489
MissForest	5.648
Amelia	6.014

04 Conclusion On the Project

To conclude (1)

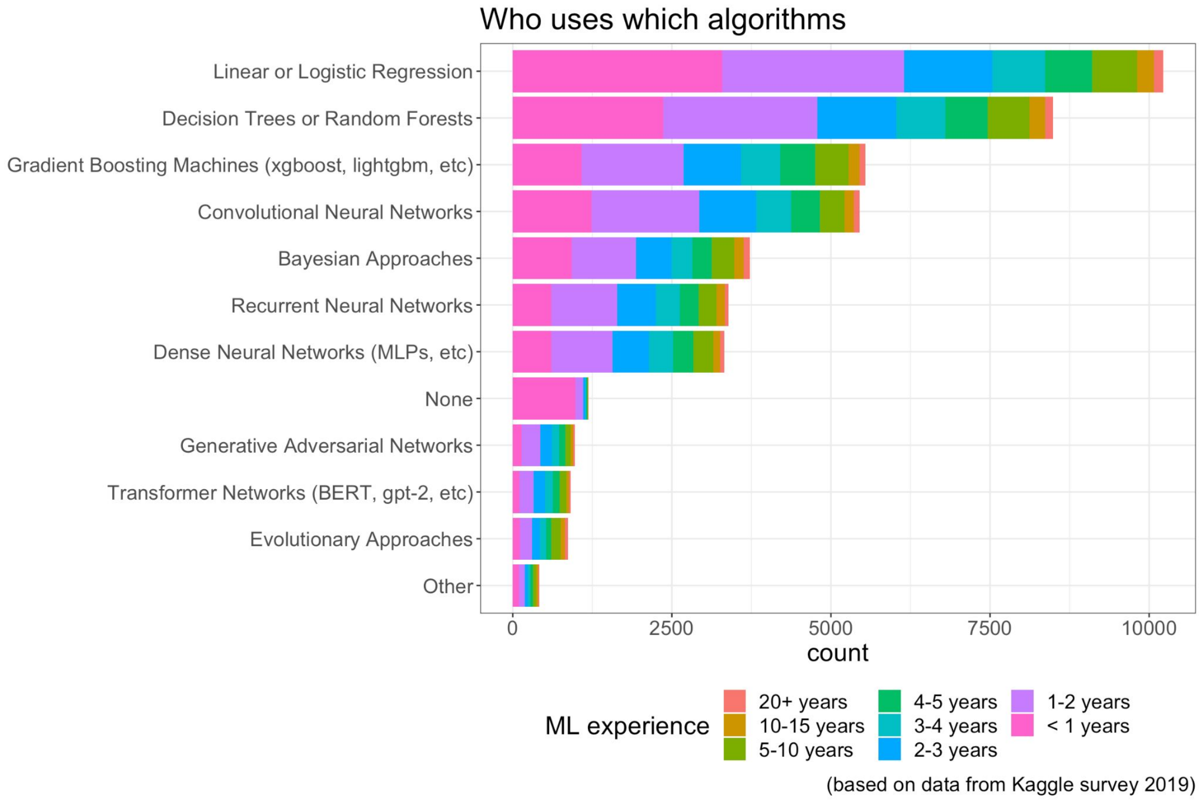
Key findings

- The use of GDP’s growth is excess in models
- Longitude is less useful than Latitude in finding common patterns
- Energy Intensity is less valuable in predicting CO_2 emission

ML models

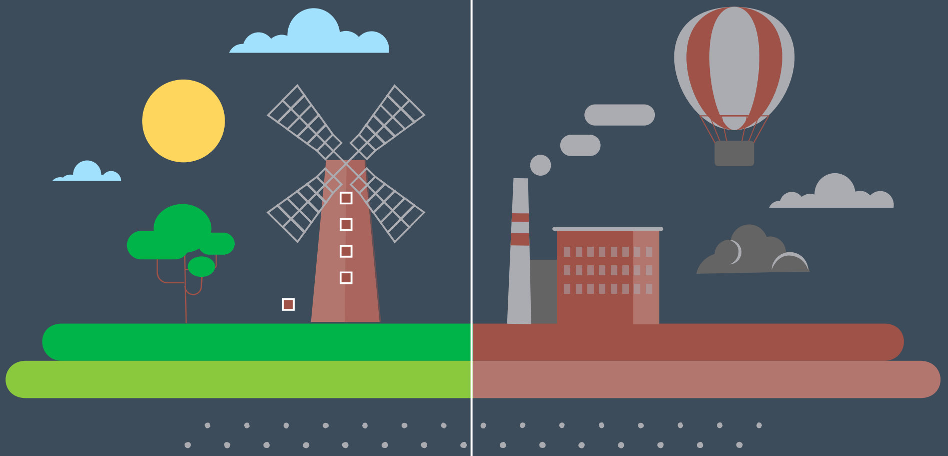
- Linear models are fast, but less likely to give exact results than other models
- Tree models, especially Random Forest give optimistic results in terms of both accuracy and time consummation
- Neural Network could be further tuned, but time consumption will goes in hand

To conclude (2)



Thanks!

Do you have any questions?



CREDITS: This presentation template was created by **[Slidesgo](http://bit.ly/2Tynxth)**, and includes icons by **Flaticon**, and infographics & images by **Freepik**