RANDOM FOREST

HUNOR TOT-BAGI

DUŠAN STAMENKOVIĆ

PMF NOVI SAD 2019

DEFINITION

- **Random Forest** is a ML algorithm which can be used for classification and for regression
- It consists of a large number of unique **Decision Trees**
- Tin Kam Ho in 1995 first used the name **Random Decision Forest**

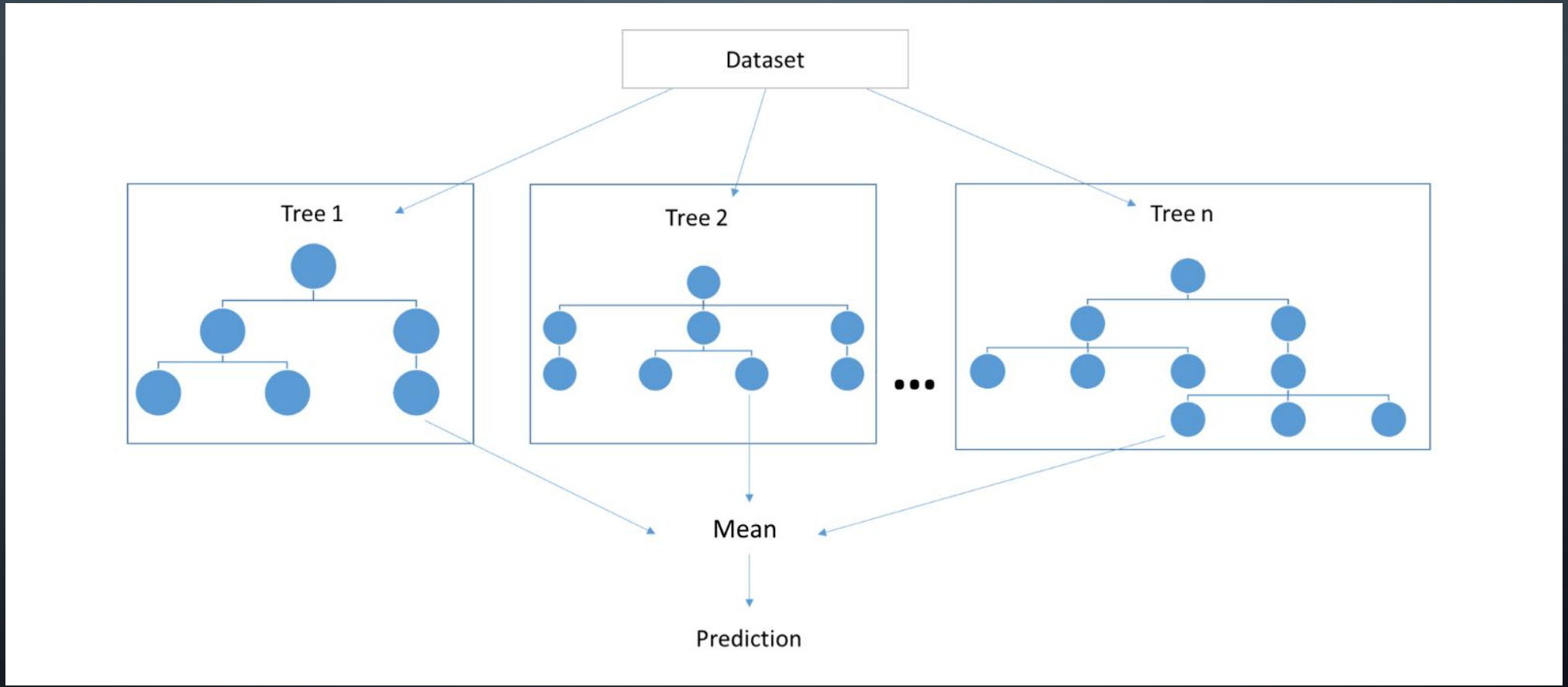
DECISION TREE

- Decision tree is the fundamental part of a RF algorithm
- Every decision tree makes its own prediction (whether for classification or regression) such that it's dividing the parameter space in disjoint rectangular regions
  - This regions have to be as "clean" as possible (only one class in region in case of classification, constant value in case of regression)
- The more decision trees we have in the RF algorithm, the more robust are predictions

THE ALGORITHM

- Every decision three is constructed in the following way:
  - Let N be the number of training samples, and M number of features
  - Import the parameters $n << N$ for number of samples and $m < M$ for number of features. Quantities n and m are constant throughout the training process
  - For each tree, randomly choose m features and n samples. Based on these the tree will make a prediction
  - Calculate which is the best split (for classification with gini-index or entropy, and for regression with mean)
  - Every tree is fully grown (splitting to the end)
- When new sample arrives, it travels through every decision tree and the prediction for that sample is the mean of all predictions (from every tree)

THE ALGORITHM



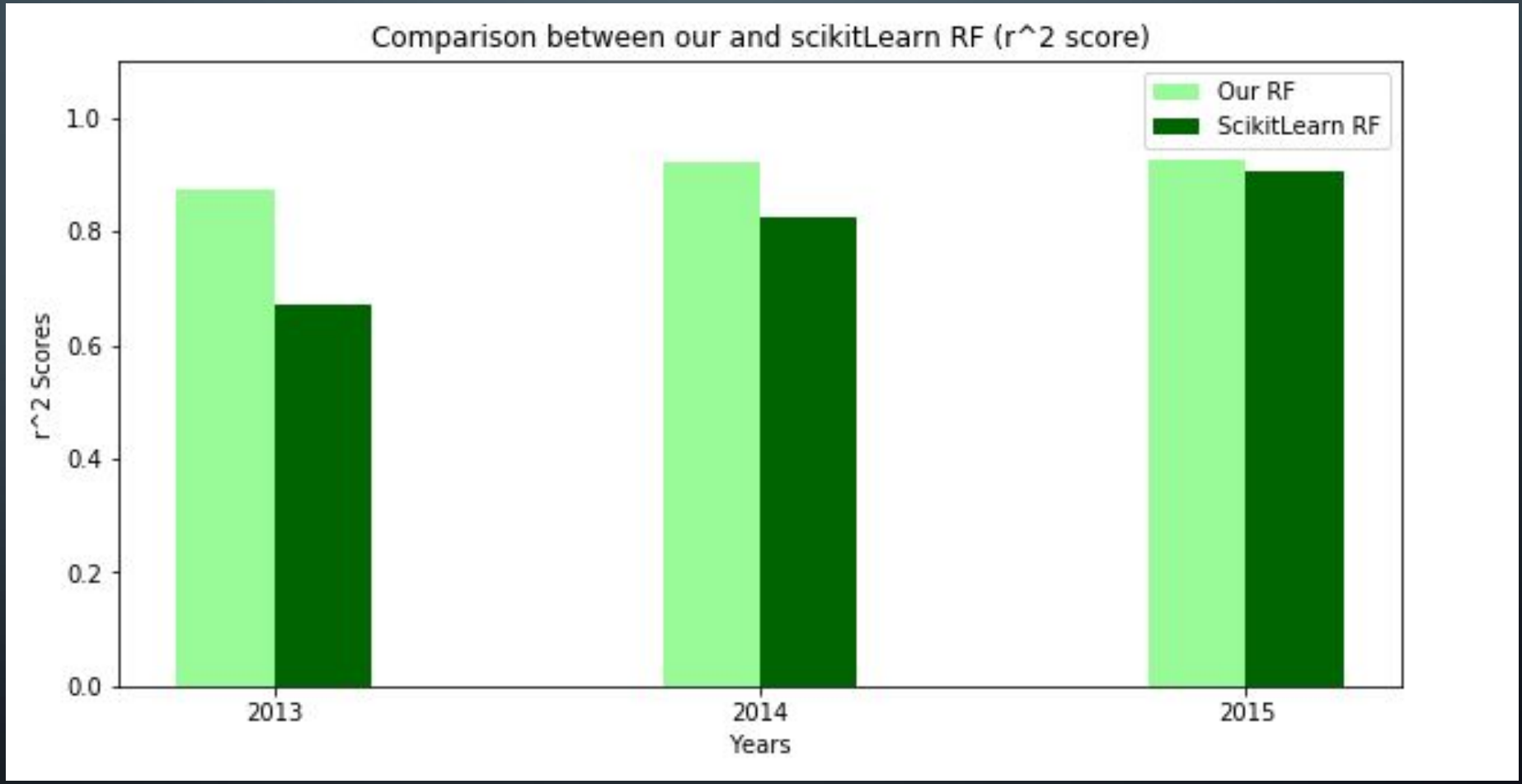
PROS

- It can be used for classification and for regression
- RF can handle well:
  - Missing values
  - Big number of dimensions
  - Large datasets
- Overfitting occurs rarely
- We can estimate which features are the most important for prediction (Feature importance)
- It's one of the most interpretable ML algorithms (less interpretable than Decision Tree)
- It handles data without preprocessing (scale or transform)

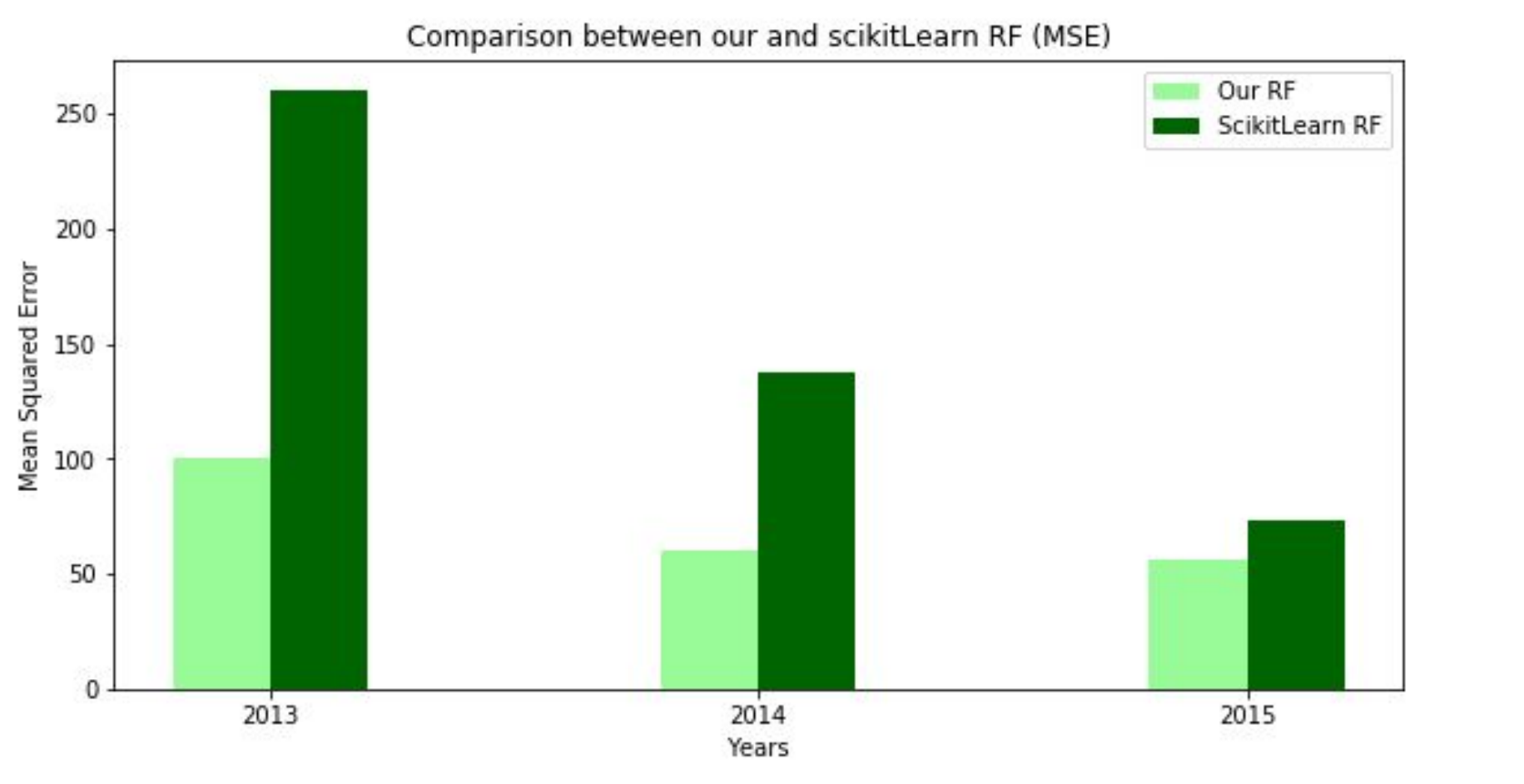
CONS

- It can overfit datasets with noisy classification/regression tasks
- Feature space is being divided into rectangular regions with sides parallel to axes

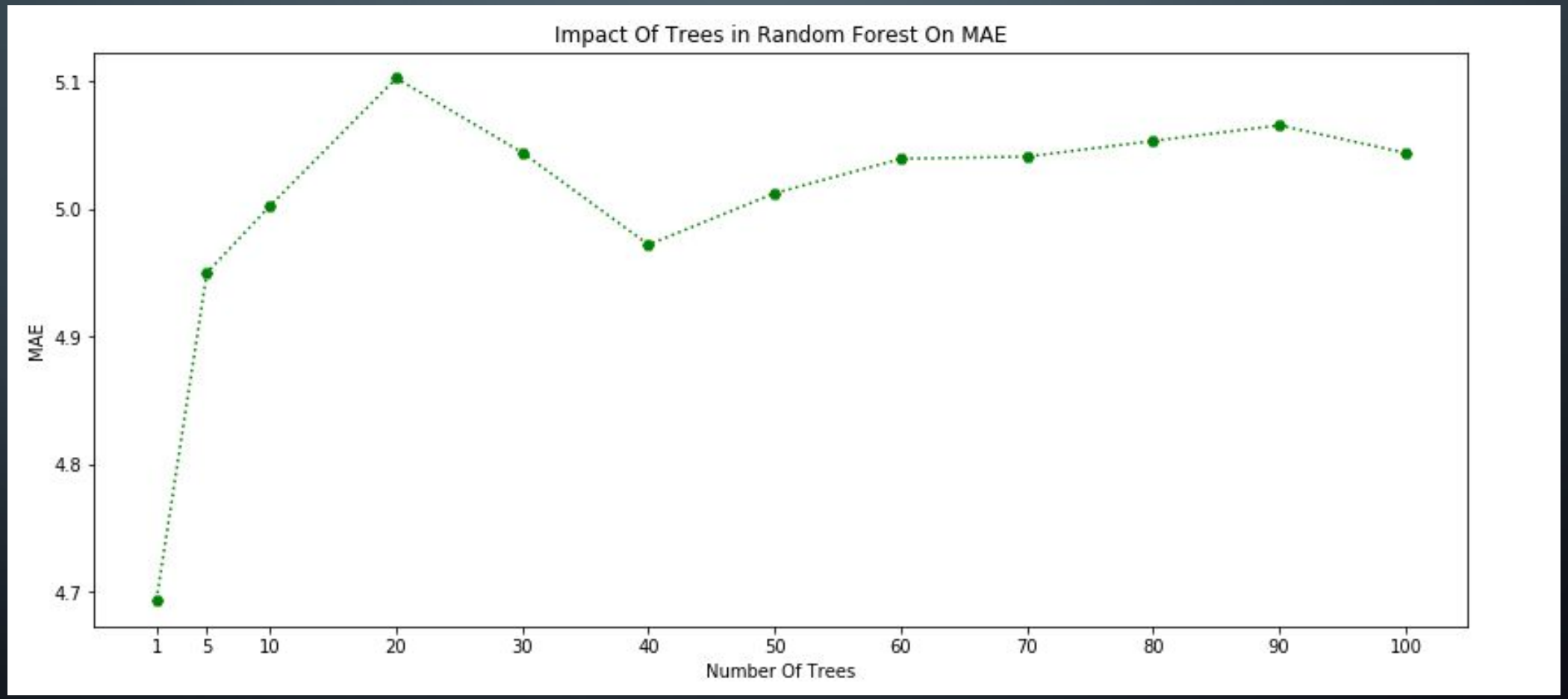
CROSS-VALIDATION AND SCIKIT-LEARN RF



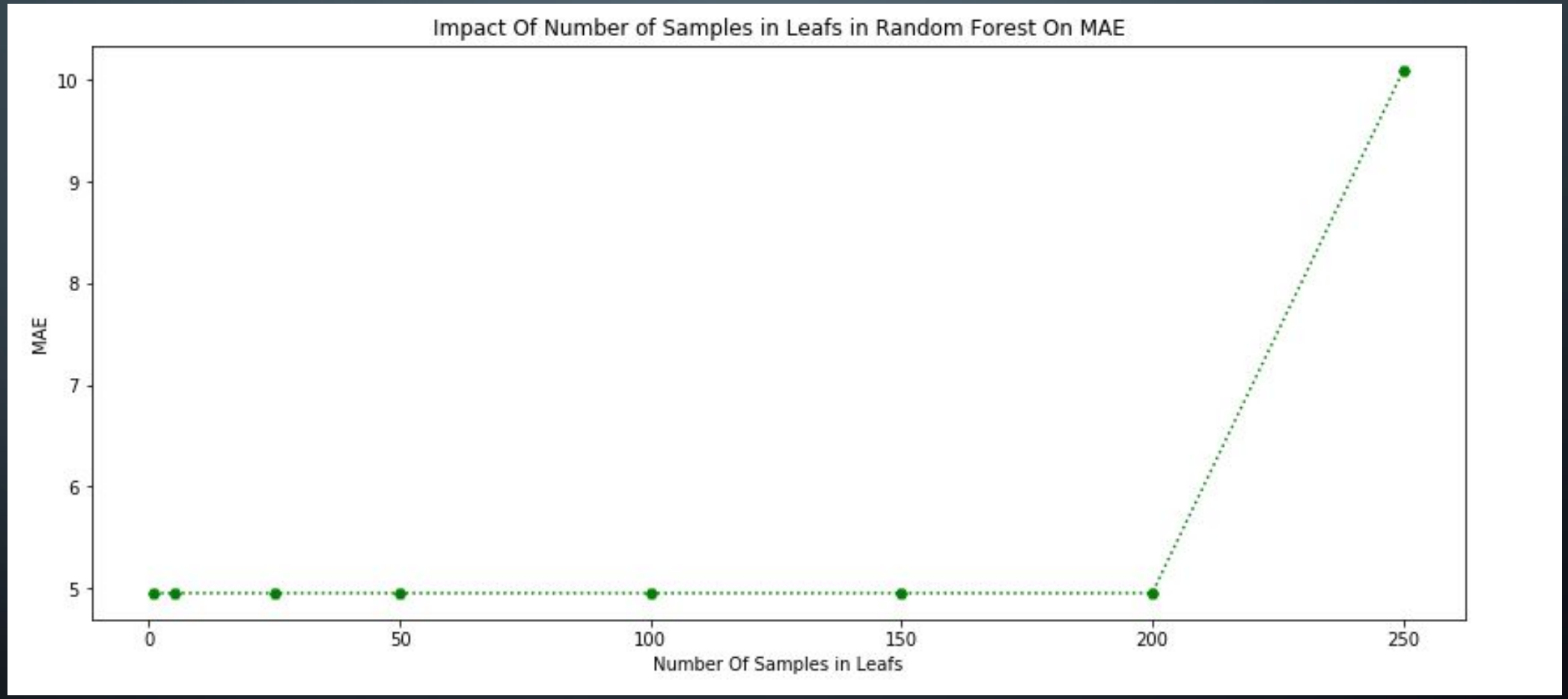
CROSS-VALIDATION AND SCIKIT-LEARN RF



ADJUSTING THE NUMBER OF TREES IN THE FOREST



ADJUSTING THE NUMBER OF SAMPLES IN THE LEAF



THANKS FOR YOUR ATTENTION! ☺️