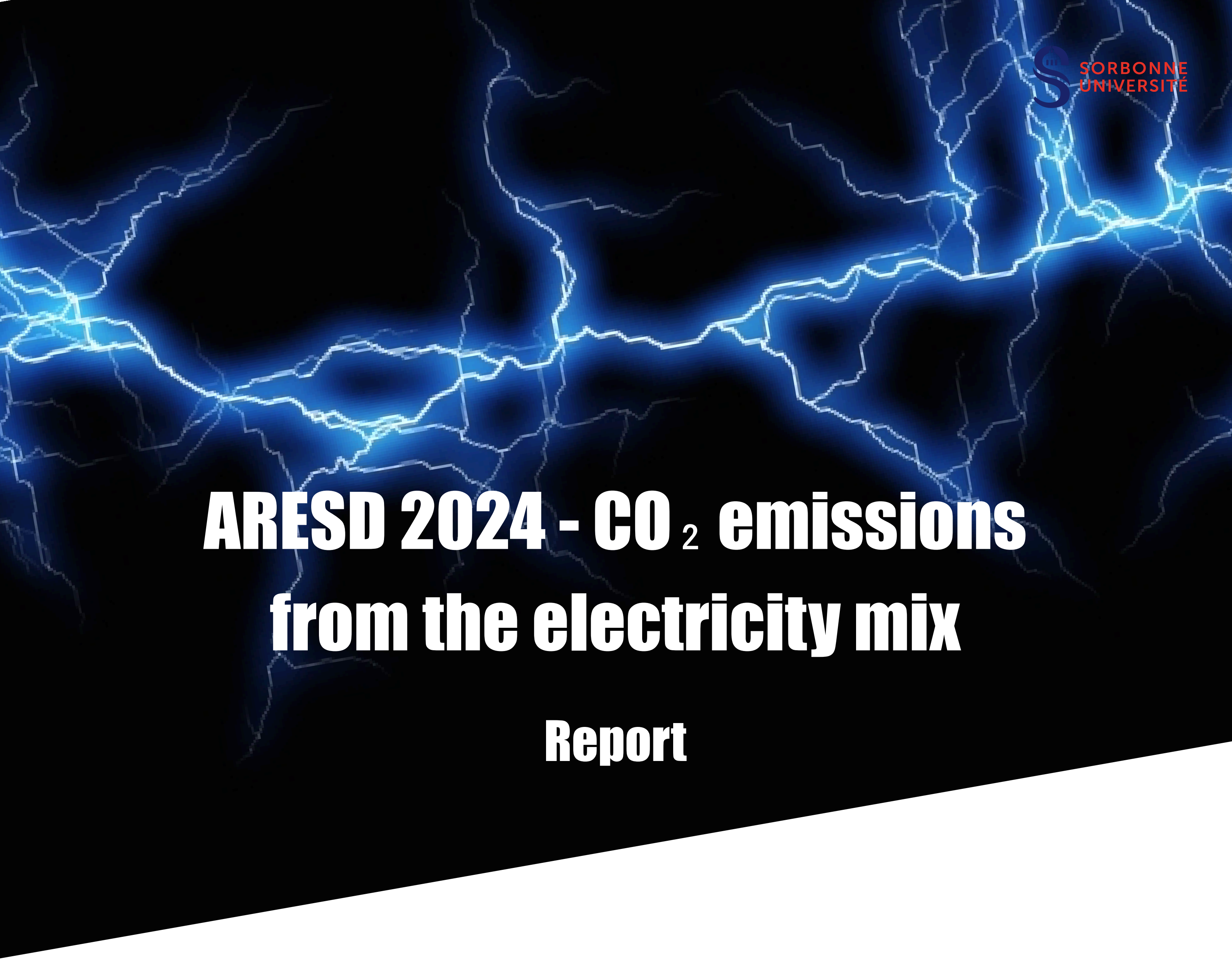


SORBONNE UNIVERSITÉ

ARESD 2024 - CO$_2$ emissions from the electricity mix

Report

YUKSEKKAYA Nehir
ALIAOUI Bochra
BRAHIMI Aymen
CAMGOZOGLU Lara

Summary

- Introduction
- Presentation of the models
  1. Bayes Classifier
  2. kNN
  3. Decision Tree
- Conclusion

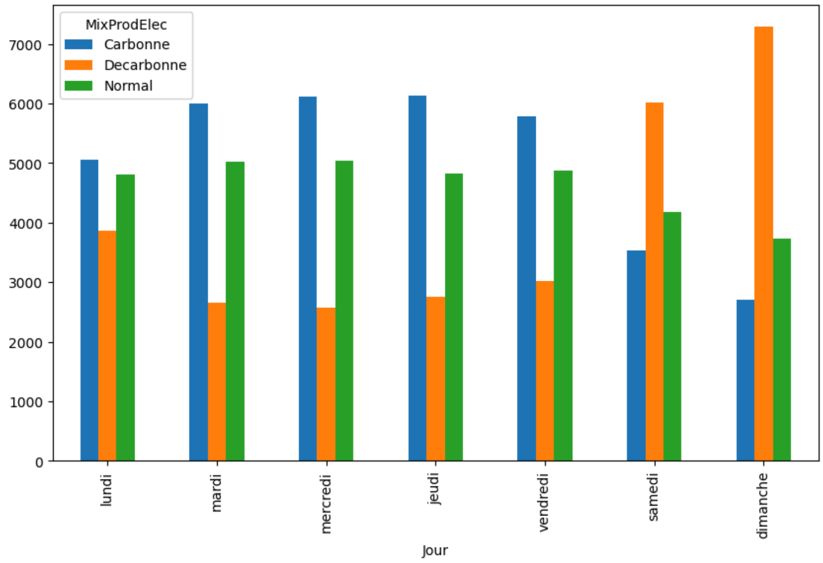
Introduction

This class's objective is to predict the carbon quality of the electricity generation mix (Carbonised, Normal, Decarbonised) given weather and calendar variables by using different models and comparing their accuracy rates to decide which one is the most reliable one.

Throughout this report, we will explore the principles of the Bayes Classifier, kNN, and Decision Tree, aiming to gain insights into their roles in modern data analysis and classification tasks.

The datasets are open source and have been provided by the French electricity transmission network (RTE) and météo France. We have a lot of variables available in the dataset such as DateTime, PositiondansAnnee, Mois, Jour… which are discreet and non binary variables. However, we also have binary variables like JourFerie, VacancesZoneA, VacancesZoneB. Finally, we also have continuous variables like Precipitation, Humidity, Nebulosity which cannot be studied here because we don't have the resources to do it. The main difficulty here is to study binary and non binary variables with the same algorithm.

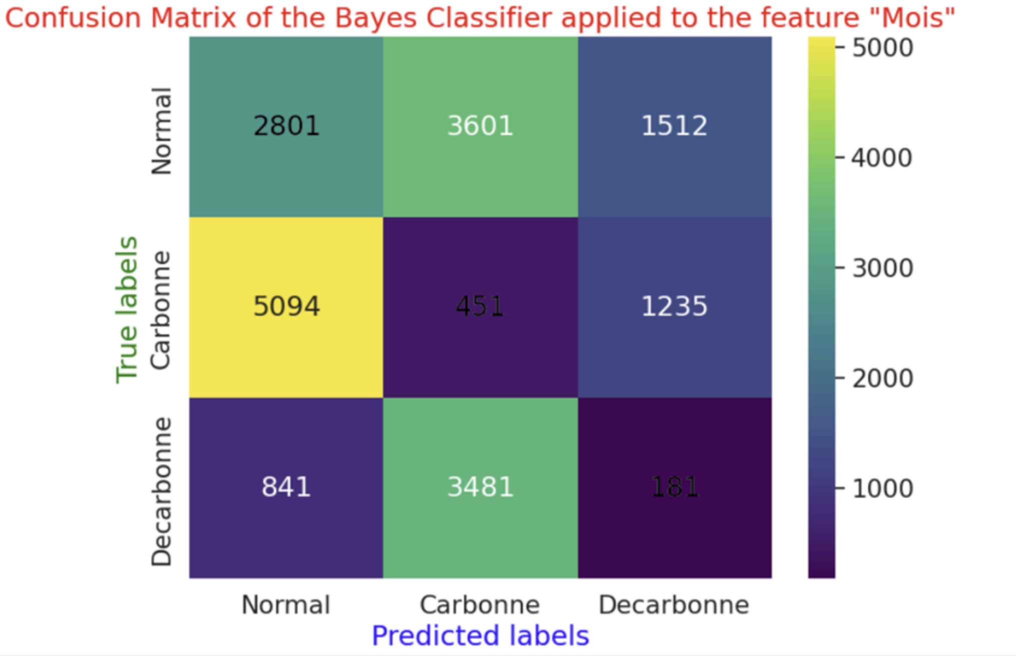
Here is an example of the predictions that a model gave us using the feature "Jour" (which we turned into a discreet variable using a label encoder):



Model 1 : The Bayes Classifier

The Bayes Classifier is a data model: it calculates the probability of each class (in our case: « Normal », « Carbonne », « Decarbonne ») using features describing the element to classify. Then, it takes the highest probability to predict the class that the element belongs to. It is also called Naïve Bayes because we assume that all features are independent. However, in our subject, the features can be somewhat related to each other which makes the model not that accurate therefore can falsify our results.

This method is a simple algorithm which means that it does not give us a reliable classification (hence the low percentage of accuracy: 30-55%). Plus, during the class, we did not study how to apply our Bayes Classifier to continuous variables which leads us to have less features to work with. Here is an example of the predictions that the Bayes Classifier gave us while applying it to the feature “Mois”:



Confusion Matrix of the Bayes Classifier applied to the feature "Mois"

Model 2 : kNN

The k-nearest neighbor algorithm, abbreviated kNN, belongs to the family of machine learning algorithms.
The k-nearest-neighbor algorithm is a supervised learning algorithm, so it requires labeled data. From a set of labeled data (in our case, it is named data), it will be possible to determine the label of a new element that does not belong to our set.

The principle of the kNN algorithm is based on proximity, but thanks to some mathematical notions like calculating the distance between points on a graph, this will make things clearer for us.
The algorithm is versatile. It can be used for classification, regression and information retrieval (as we'll see in our case).

Its principle to reach our goal is playful:

1- Load data

2- Initialize k to the selected number of nearest neighbors

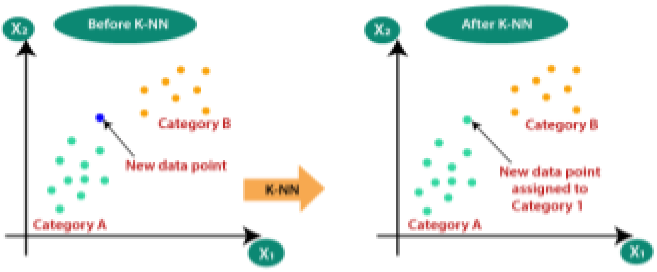
3- For each example in the data: calculate the distance between our query and the loop's current iterative observation from the data.

4- Sort this ordered collection containing distances and indices from smallest distance to largest (in ascending order).

5- Select the first k entries in the sorted data collection (equivalent to the k nearest neighbors).

6- Obtain the labels of the k selected entries

8- Return the most frequent value of k labels



https://static.javatpoint.com/tutorial/machine-learning/images/k-nearest-neighbor-algorithm-for-machine-learning2.png

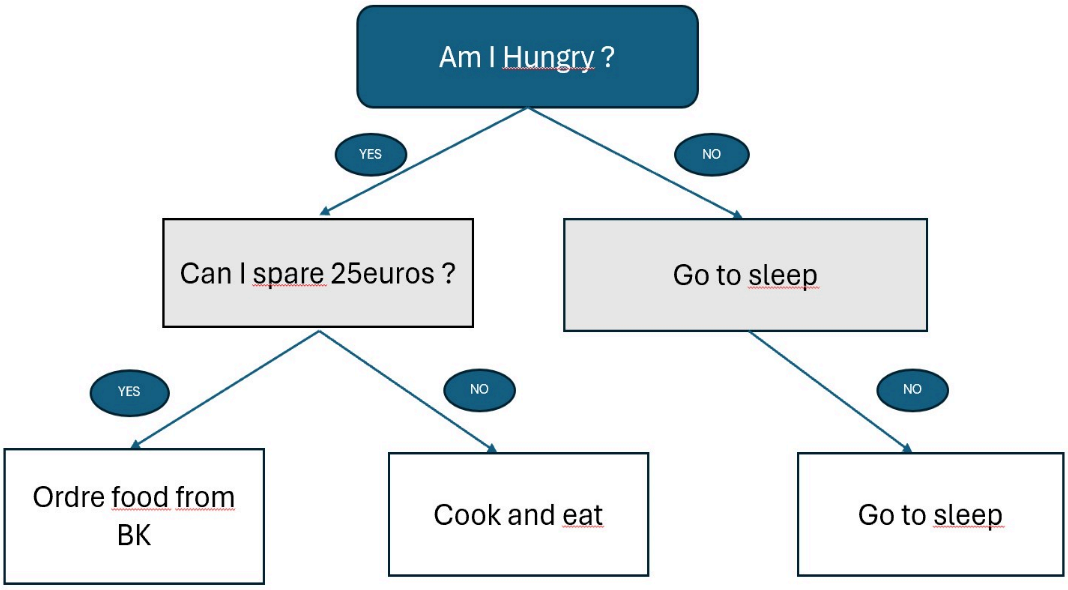
Model 3 : Decision Tree

The general aim of a decision tree is to explain a value from a series of discrete or continuous variables. This is the classic case of a matrix X with m observations and n variables, associated with a vector Y to be explained. The values of Y can be of two kinds:

- Continuous: this is called a regression tree
- Qualitative: we call this a classification tree

These methods have a number of advantages: they are fairly efficient and non-parametric. In principle, they will partition the objects of our study, producing groups that are as homogeneous as possible from the point of view of the variable to be predicted, taking into account a hierarchy of the predictive capacity of the variables considered.

Let's understand a decision tree from an example: Yesterday, I skipped dinner because I was busy. Later in the night, I felt hungry. I could have gone to sleep as it is but as that was not the case, I decided to eat something. I had two options, to order something from outside or cook for myself.
I figured if I order, I will have to spare 25 euros on it. I finally decided to order it anyway as it was pretty late and I was in no mood of cooking. This complete incident can be graphically represented as shown here:



Conclusion

In conclusion, the Bayes Classifier, K-Nearest Neighbors (KNN), and Decision Trees emerge as essential components in the realm of data analysis and machine learning, each offering distinct methods for classification tasks.
Thanks to the Average Class Accuracy that we defined, we could obtain accuracy rates for each model: that way we could decide which one is the most efficient and accurate one. The Bayes Classifier is the one with the lowest accuracy rate, around 30 to 55% of precision depending on the feature. The two other models (kNN and Decision Tree) have a higher accuracy rate starting from 45% and can reach higher accuracy rates but due to some submission problems, we couldn't evaluate those values.

Even though the algorithms are different, we could notice that they approximately have the same behavior for a certain case. They usually predict correctly the class “NORMAL” but they get lost trying to distinguish the other classes “CARBONNE” and “DECARBONNE”. That may be because some of the observations of these classes are similar : except for the seasons, the calendar variables can’t really help us predict if the electricity mix is carbonised or decarbonised.

Given the volume of data to analyze, classification models produced very long execution times due to the big amount of data at our disposal. The fact that we couldn't analyze all the data may have had a negative impact on the accuracy of our models, as we could end up with a dataset that was not diversified enough and did not allow for correct classification of the entire public test dataset on Kaggle.