Data Science and Machine Learning
Project UNIL_TUDOR

Goal:

In this project, our aim was to create a startup that would revolutionize the way people learn and improve their command of a foreign language.
Our main goal was therefore to predict the level of difficulty of French text. To achieve this, we had to create a model for English speakers that could classify French sentences by language proficiency levels (A1, A2, B1, B2, C1, C2).

To do this, we began by concentrating on the models we had mainly seen in class, and which had been suggested to us, such as:

- kNN model
- Logistic regression,
- Decision tree classifier
- Random forests classifier.

After training these models, we realized that the results obtained didn't really match our expectations. That's why we set out to find a new model that would perfectly match what we were looking for and focused on a new model: the CamemBERT model. CamemBERT is a natural language processing (NLP) model based on BERT, but designed specifically for understanding French texts, which is exactly what we need for this project.

Part 1:

For this first part, we decided to proceed in the same way for all the models. In fact, in each model we decided to use a TFIDF (Term Frequency-Inverse Document Frequency) vectorization approach because this model converts text into numerical vectors, which is necessary because the machine learning models, we use cannot work directly with plain text. After that, we assembled the vectorizer and classified it in a pipeline to finally train the model.

Once we had all this, we decided to show the Accuracy, Precision, Recall and F1 score of the model. For each model we realized that the scores obtained were not very satisfactory, so we tried to improve the score by performing hyperparameter optimization with GridSearchCV. We can see that this worked very well, as the score obtained after this was better than before for each model, except for the Random Forests classifier where the results obtained after hyperparameter optimization were less good.

Part 2: