This project is the follow-up of the previous project that you applied named entity recognition (NER) on medical text on Medical Transcriptions [1]. Initially, remove the transcriptions having category labels less than 50 in the corpus as in [2]. Apply data preprocessing techniques also following the steps in [2]. For feature extraction, apply Bag-of-Words (CountVectorizer) and TF-IDF (TfidfVectorizer) separately. Implement Multinomial Naïve Bayes, Random Forest, XGBoost, LightGBM for the traditional machine learning algorithms of the medical text classification process. Then, apply at least one complex deep neural network architecture (ensemble learning) using 1D CNN, LSTM and GRU. Show the confusion matrix, accuracy, precision, recall and F1-score for each category class of the implemented solutions.

In the next phase, use the NER code previously implemented in the first part of the project. Use the labeled named entities and their category labels as the input, then follow the same training and evaluation steps.

Finally, apply SMOTE oversampling method [2] for the best accuracy values in the previous two phases and compare accuracy, precision, recall and F1-score with and without oversampling. Write a report that explains and illustrates the results step by step. Upload the source code and the report to ADUZEM.

[1] https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions/data

[2] https://www.kaggle.com/code/ritheshsreenivasan/clinical-text-classification