**Project proposal – Active Learning featuring GNNs**

**Team members:**

* **Saleem kherladen,** [saleemkheralden21@gmail.com](mailto:saleemkheralden21@gmail.com)
* **Majd**
* **Lana**

**Project Title:** Active Learning featuring GNNs

**Subject:** integrate GNNs in AL scheme on tabular data

In this project we’ll explore the effect of GNNs on the AL scheme and whether it improves the performance of the classification model.

Our plan to explore this topic is the following, we’ll collect multiple datasets, while focusing on how much the labeled data is clustered or not, our intuition is that data that somewhat clustered would be better for GNNs that non-clustered data, our scheme would be the following:

1. Given a tabular dataset, we’ll convert it to graph, such that each datapoint is a vertex in the graph and the connections between vertices would be constructed using KNNGraph method, Cosine similarity or another distance metric (we’ll explore the effect of each edge creation method on the performance of the pipeline)
   1. After creating the graph, we’ll explore the effect of different Uncertainty metrics (entropy, margin-based, query-by-committee, …) and we’ll explore another centrality metrics as uncertainty metrics, such as PageRank score, degree…, those metrics would be the score we’ll give to the datapoints in order to add them to the labeled set.
   2. In parallel, the graph of the data will be passed through a GNN, in order to improve the embedding of the data to be classified, we’ll dive deeper into how to train the GNN model, and whether we’ll train it on the picked datapoints, or to include another datapoints which are not picked (this will be explored in the project)
2. After selecting the datapoints and getting their labels from the oracle, and after getting the new embedding of them, the data will be passed through the classifier, in order to train it, like the classic scheme of AL.

In conclusion, the topics we’ll explore in this project:

1. Whether converting tabular data into graph improves the AL scheme.
2. Comparison between edge creation methods (using KNNGraph method, Cosine similarity…)
3. Comparison between uncertainty metrics (entropy, margin-based, query-by-committee) with the proposed metrics (PR, degree…)
4. Methods to train the GNN in this scheme.