MAIS 202 - PROJECT DELIVERABLE 1

ll use the wine review dataset on Kaggle. This dataset contains more than 130 000 reviews of wines in the form of a description of the taste, as well as the country, region, points and price. I chose it because it is large and has a lot of different features that I could use to feed my machine learning model.

I will have to preprocess the data. First, I will trim the data to keep only the columns of interest, namely the description and the price. Then, I will have to separate the text data to keep only relevant words. I will remove the noise (comas, dots) and stopwords. I will normalize the data to remove capital letters. It might be necessary to lemmatize the words, but this will be evaluated as the project progresses. Using a word embedding tool, I will embed the words for each review in a vector space. For the price, I will partition the prices of the wines into different categories. Each category will represent a price interval and will be given a number. For instance, the price range $0<p<$10 might be given the value of 0.

By partitioning the prices into different categories, I have transformed the problem into a classification problem. With the word vector, I will use a support vector machine to draw a boundary between the price different prices based on the input. Then, it could be possible to also use a multi-label random forest. I will also try using a neural network to see if it gives better prediction. If I have the time, the same dataset and a variance of the model could be used to forecast other features, like the region.

Finally, I will present my project in a web app. The user will input a description of the wine and the model will try to determine the price category.