# Machine Learning Report Project

**FOZAME ENDEZOUMOU Armand Bryan$^{\mathbf{PGE4}}$, Maheni SOUMAH$^{\mathbf{PGE3}}$, Jessica MBOUNKAP$^{\mathbf{PGE3}}$, Habiba DJIGO$^{\mathbf{PGE3}}$, Darryl TOWA$^{\mathbf{PGE3}}$, LEUMALEU MBOUYOM Arnold$^{\mathbf{PGE4}}$**

**Abstract**

This academic project, supervised by Dr. Hanoune, aimed to master fundamental concepts in machine learning. It involved completing four case studies, along with a few additional tasks. The first project focused on classification and regression, the second on image classification, the third on unsupervised learning, and the last on object detection and segmentation. Additionally, participants were tasked with implementing classification and regression using AdaBoost, as well as developing a convolution algorithm from scratch, including stride and padding. These projects were completed and documented on our GitHub repository, which can be found at the following link: `https://github.com/Bryan-Foxy/ml-project`. In the remainder of this report, we will explain our choice of algorithms used and present the results obtained.

## 1 Introduction

To successfully carry out our projects, the first critical step was to carefully select the datasets we would work on. After thorough consideration, we chose a variety of datasets, each targeting a specific aspect of machine learning. These datasets are stored in a shared Drive folder, accessible via the following link. Here is an overview of the selected datasets that will form the foundation of our work:

- For classification/regression : We opted for the renowned Wine dataset, which offers a fertile ground for various predictive analyses.
- In image classification : We focused on images from American Sign Language, a choice motivated by the desire to make technology more inclusive.
- For unsupervised learning : We chose the dataset of cancers in Lake County, Illinois, USA, to explore models capable of identifying patterns and anomalies.
- Regarding image segmentation : Our attention turned to images and masks of lungs affected by Covid-19, a timely and critically important subject for public health.

To facilitate collaboration and share our progress, we created a GitHub repository, a central space where our work can be versioned and reviewed efficiently by all team members. Additionally, to leverage advanced computational resources, we chose to use Google Colab, which allows us access to powerful GPUs, significantly speeding up the processing time of our analyses.

This infrastructure enabled us to dive into exploring these datasets with all the necessary resources at our disposal. Our goal is not only to extract relevant insights from these data but also to develop robust and efficient machine learning models. Each project, targeting a different aspect of machine learning, is designed to deepen our theoretical understanding while confronting us with practical challenges, thus preparing us to tackle real-world issues.

## 2 Classification/Regression

In our inaugural project, we embarked on both classification and regression analyses using the Wine dataset. A key aspect of this project was not only to analyze this dataset but also to integrate it with another dataset from the same domain, albeit with differing columns. This task required a meticulous approach to data preparation and analysis.

During our initial analysis, we identified missing data within our datasets. To address this issue, we employed the Newton Interpolation method, a strategic choice for estimating missing values by leveraging nearby data points. The Newton Interpolation formula we applied is as follows:

$$P_n(x) = \sum_{i=0}^{n} \beta_i w_i(x) with w_i(x) = \prod_{j=0}^{i-1} (x - x_j), \forall i \in [1, .., n] and w_0(x) = 1$$

This equation succinctly captures the essence of Newton's Interpolation, where $P_n(x)$ represents the polynomial of degree $n$ approximating the function. Each $\beta_i$ is a coefficient determined through divided differences based on the known data points, and $w_i(x)$ are the basis polynomials constructed from the input data points $x_j$.

After addressing the missing data and thoroughly analyzing the datasets, we proceeded with our experiments, which encompassed four distinct analyses: two focused on classification/regression with a single dataset, and two more utilizing the merged dataset. For these tasks, we chose the XGBoost model, renowned for its effectiveness and efficiency across various machine learning challenges. These are the results:

Table 1: Table of accuracy of the both classification model

| Models | Accuracy |
|---|---|
| Model$_{classification}$ | 99.23 |
| Model$_{classification} assembly$ | 99.54 |

Table 2: Table of accuracy of the both regression model

| Models | MSE | RMSE | MAE |
|---|---|---|---|
| Model$_{regression}$ | 0.39 | 0.63 | 0.45 |
| Model$_{regression} assembly$ | 0.41 | 0.64 | 0.49 |

The use of XGBoost was pivotal, given its proven track record in handling both classification and regression tasks with high accuracy. This model's capacity to manage sparse data and its versatility in dealing with various types of predictive modeling tasks made it an ideal choice for our project.

## 3 Classification des images de langage des signe

## 4 Unsupervised learning

In this section, we have applied our unsupervised learning skills with the goal of visualizing the different cancer regions across Lake County. Additionally, we aim to understand how dimensionality reduction works by implementing it.

Initially, we are working with two distinct types of data:

- A .geopandas file that contains the geographical coordinates of Lake County.
- A .csv file that includes data on cancer cases found in Lake County.

To display the geographical file, we utilized `geopandas` for mapping visualization. After preprocessing the .csv data, we determined the number of clusters using the elbow method and then applied KMEANS to identify distinct regions. These regions are then overlaid on our map. Also to find the best k, we have perform the famous elbow method and the Figure **??** is the results.

Next, we move on to dimensionality reduction to determine the number of components necessary to retain 95 percent or more of the information. This technique is employed for feature selection as some columns may contain white noise without contributing any useful information. We visualize the data using PCA and t-SNE in 2D and employ PCA in 3D, which allows us to clearly see the clusters identified earlier with KMEANS.

The results are in the Figure **??**

## 5 Segmentation des images X-Ray des poumons atteint de Covid19

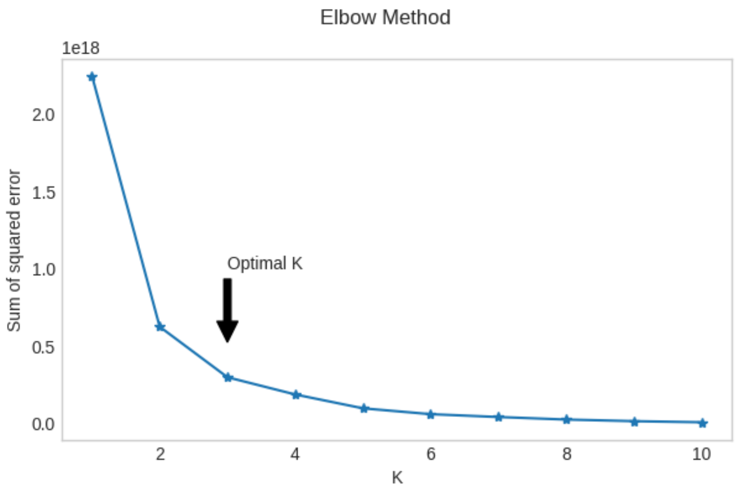
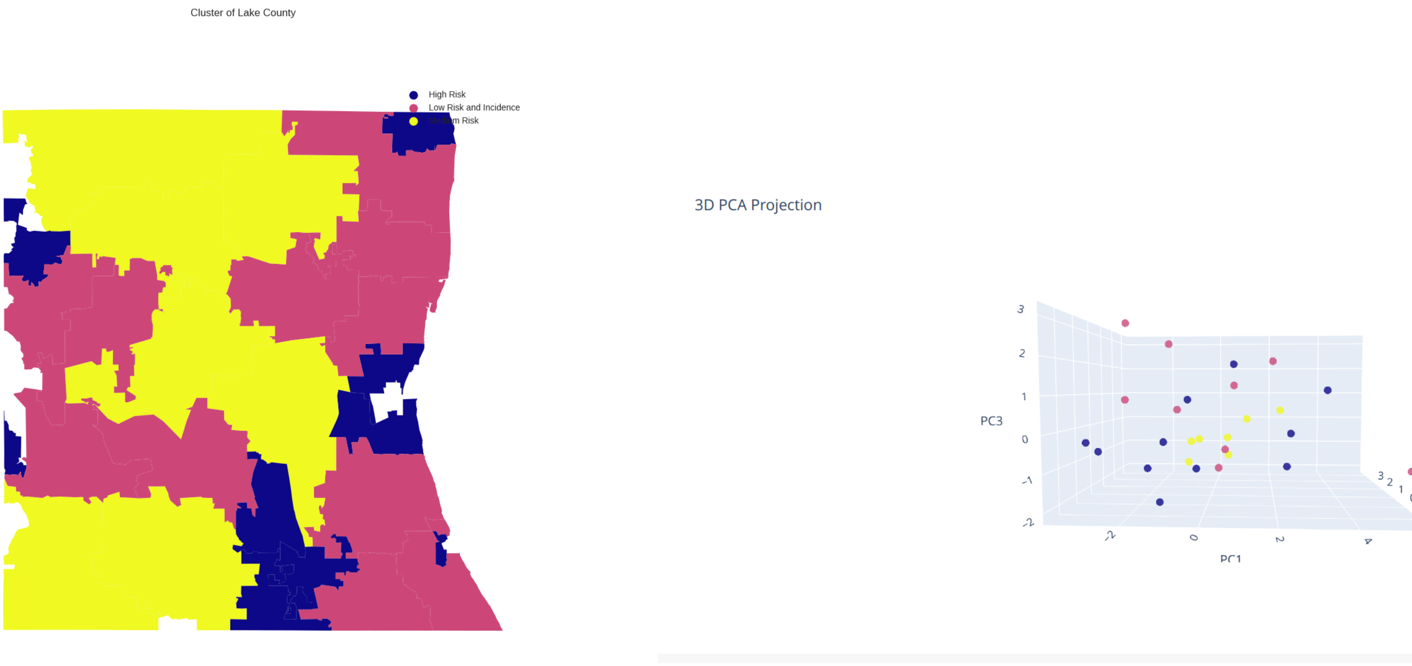


Figure 1: Optimal K



(a)

(b)

Figure 2: Visualization of KMEANS on the map and PCA 3D