Proposal (Group 14): Deriving Insights from an Obesity Dataset

Problem:
The dataset from UCI is particularly built for classifying 7 types of obesity levels based on simple features. We would mainly like to delve into the EDA and classification and clustering of the data to predict the type of obesity.The dataset also contains regional information and so we would like to also estimate the obesity levels based on regional nutritional values.

Dataset:
UCI repository:
https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

This dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records.

Proposed Solution and Real world Application :
Our first step would be to conduct an exploratory study to know the characteristics of the participants in the dataset and know their eating habits and physical activity using simple techniques such as finding the correlation between the features, deriving results such as BMI and calorie related metrics along with some geospatial insight as it specifically contains data from 3 regions.
After exploring the data we'll step into the classification and segmentation step to either go for a binary classification or multivariate classification for the obesity levels based on the features and other computed parameters. Different types of models using machine learning techniques such as logistic regression, adaboost, etc will be used to find the optimal solution.
Coming to the real world application of this research, it can be used to build software tools in healthcare systems that require an insight on the patient's well being. Another application would be to build regional insights provided we have a lot of data for the world for census related applications. Another would be a recommender system which gives out the meal plan for a certain kind of test subject having a certain condition. We would delve into one of the applications mentioned here by using streamlit.

(Please refer the second page for the associated steps - Also note a particular person's name is not mentioned in each of the steps as we're still deciding as to who will be best fit for the tasks, Later a task related chart would be shared or put in the ppt)

Step	Estimated completion time	Person(s) in charge
Data Cleaning and Preprocessing	1 week	2 people
EDA	1 week (Can be carried out simultaneously with the first step)	2-3 people
Building a classifier	2 weeks	2-3 people
Building a visualizer or app using Streamlit	~1-2 weeks	2 people
Powerpoint Presentation	1 week	1-2 people