**Medical Statistics Analysis**

**Name Student ID**

Umair Imran L22-8370

Bilal Ahmad L22-

Muhammad Shahzad L22-

Najam Ul Islam L22-

**Problem and Data Set Description**

In analyzing country-wise data, a significant challenge arises in conducting collective analysis to determine which country requires specific resources. While individual country analyses provide valuable insights, performing a comprehensive analysis on a collective scale proves difficult.

This project addresses the challenge by predicting population growth rates and life expectancy on a larger scale. Using a drill-up approach, we will begin with country-level data and progress to continent and region-level analysis. The dataset spans from 2000 to 2024 and includes dimensions such as population, life expectancy, and trends in diseases across various countries.

The primary objective of this project is to offer a unified solution for collective analysis of health and resource needs across countries. It will also enable predictions of specific variables for individual countries or regions.

The dataset includes key dimensions such as population, life expectancy, trends of diseases over time, causes of mortality, and health expectancy rates for both genders. We will employ unsupervised learning techniques to effectively analyze and predict these factors.

# Preliminary ideas on how you plan to address it (models/algorithms/techniques)

Since the features of different nodes can have different lengths in the dataset, the first step of this work is to construct another feature representation of each node that has the same length (dimension) that can facilitate the node embedding process afterward. After analyzing the features, we note that each feature is a subset of the set of integers [0:4044]. Therefore, a direct approach to constructing a new feature vector is to construct a 4005-dimensional vector with elements 0/1 representing the existence of a corresponding element. This will generate a 4005xNv feature matrix. We may then consider using SVD to reduce the dimension of the feature matrix.

After obtaining the feature matrixwe will explore multiple machine-learning algorithms which can use structural information for node classification. One of the famous algorithms for node classification in graphs is Graph neural networks. We will explore Graph neural networks for solving our problem. For this purpose, we plan to use specialized TensorFlow-based libraries such as Spectral, StellarGraph, and GraphNets. Additionally, we also plan to explore Graph Convolution networks. The GCN model is based on the graph convolution layer. This layer is like a dense convolution layer that incorporates the adjacency matrix of the graph to use information about the connections of nodes. We will also study and use Graph ATtention Network (GAT) for solving classification problem [2]. Besides this, we will explore GraphSAGE techniques, Inductive GraphSAGE and directive GraphSAGE.

One of the common problems while working on a graph-based dataset is that traditional machine learning algorithms use numerical data. Hence, transforming the structural information of a network into numerical representation is an important task. In this situation, the Node2Vec algorithm helps to translate the nodes of a Graph to an embedding space. This algorithm preserves the structural information. Node2Vec is a famous algorithm for classification problems, and we also plan to explore it [3].

**Software Tools**

We plan to use Python as our primary programming language. Below is a list of specific libraries and tools that will be employed throughout the project:

* **GitHub**: For efficient project management and version control.
* **Selenium**: As a data scraping tool for gathering required information.
* **Power BI**: For performing data visualization tasks.
* **Matplotlib**: To assist with data exploration and graphical analysis.
* **Pandas** and **NumPy**: For data manipulation, processing, and numerical computations.
* **Sklearn**: For implementing unsupervised learning algorithms and statistical modeling.

**Expected Results and Evaluation**

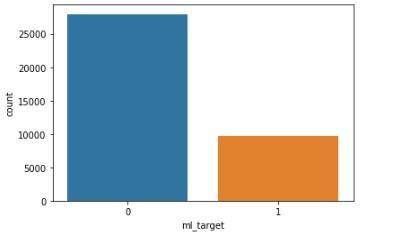
This project primarily focuses on regression tasks, supplemented by some classification tasks. We anticipate obtaining regression results that reflect various health and demographic statistics across all countries. For the classification component, we expect to classify countries into categories such as low income or high income, as well as to categorize countries based on specific disease prevalence.

To evaluate our regression results, we will use metrics including Mean Squared Error (MSE) and Root Squared Error (RSE). For the classification tasks, we plan to assess our results using several key performance metrics, including accuracy, sensitivity, specificity, and F1 scores. These evaluations will provide insights into the model's performance and its effectiveness in predicting health-related outcomes and demographic classifications.

**Preliminary Results and Data set explored.**

For this project, we have explored the data set used by Benedek Rozemberczki et.al [4] in their research paper related to Multi-scale Attributed Node Embedding. This dataset is based on a homogenous graph where nodes represent the subset of people who use GitHub. The dataset has 37,700 nodes and 289,003 edges. The transitivity of the graph is 0.013 and the density is 0.01. The nodes are binary labeled 0 and 1. The Ids of nodes and the names of users corresponding to different Id numbers are also mentioned.

We performed statistical analysis of data. The targets in the 25th and 50th percentile is 0 and in the 75th percentile are 1. To be exact 27961 targets are denoted as 0 while 9739 are denoted as 1. Hence, this is clearly a class imbalance problem. That is why we will use an evaluation matrix that is suitable for class imbalance problems i.e., sensitivity and specificity.



After analyzing the features for each node, we realized that each feature is a subset of the set of integers [0:4044]. We plan to find an alternative representation of features.

**Outline of the Work-to-Do**

1. **Dividing the Modules**
   * Assigning tasks and dividing the modules among all team members on GitHub.
2. **Data Scraping**
   * Scraping data from the major website (WHO) for all countries.
3. **Exploratory Data Analysis (EDA)**
   * Performing exploratory data analysis to understand the dataset.
   * **Major Techniques**: [e.g., Summary Statistics, Data Visualization, Correlation Analysis, Distribution Analysis]
4. **Feature Engineering**
   * Implementing feature engineering based on the following techniques:
   * **Major Techniques**: [e.g., Normalization, One-Hot Encoding, Binning, Polynomial Features]
5. **Data Visualization**
   * Using Power BI for data visualization to present insights effectively.
6. **Machine Learning Models**
   * Applying machine learning models to make predictions based on the data.
7. **Model Evaluation**
   * Conducting an iterative evaluation of models to refine performance.
8. **Application Development**
   * Converting the project into a real-world application using Streamlit and Flask (tentative).
9. **Chatbot Development**
   * Developing a chatbot that can retrieve data from the WHO in real time based on user queries for better analysis (tentative).

# Reference

1. Benedekrozemberczki/datasets: A repository of pretty cool datasets that I collected

for

Network Science and Machine Learning Research., GitHub. https://github.com/benedekrozemberczki/datasets#github-social-network (Accessed: February 12, 2023).

1. Veličković, P. et al. (2018) Graph attention networks, arXiv.org. Available at: https://arxiv.org/abs/1710.10903 (Accessed: February 12, 2023).
2. Scalable feature learning for networks, node2vec. Available at: https://snap.stanford.edu/node2vec/ (Accessed: February 12, 2023).
3. Rozemberczki, Benedek, Carl Allen, and Rik Sarkar. "Multi-scale attributed node embedding." Journal of Complex Networks 9.2 (2021): cnab014.
4. Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
5. Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
6. Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103(23), 8577–82. https://doi.org/10.1073/pnas.0601602103'
7. M. Fielder, “A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory,” Czechoslovak Mathematical Journal, vol. 25, no. 4, pp. 619– 633, 1975.
8. L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 11, no. 9, pp. 1074-1085, Sept. 1992, doi: 10.1109/43.159993.
9. Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888905, Aug. 2000, doi: 10.1109/34.868688.
10. Chunaev, Petr. "Community detection in node-attributed social networks: a survey." Computer Science Review 37 (2020): 100286.
11. Ahmed, Amr, et al. "Distributed large-scale natural graph factorization." Proceedings of the 22nd international conference on World Wide Web. 2013.
12. Yang, H., Pan, S., Zhang, P., Chen, L., Lian, D. & Zhang, C. (2018) Binarized Attributed Network Embedding. In IEEE International Conference on Data Mining.
13. Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
14. Wang, Daixin, Peng Cui, and Wenwu Zhu. "Structural deep network embedding." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
15. Cao, Shaosheng, Wei Lu, and Qiongkai Xu. "Deep neural networks for learning graph representations." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.
16. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K. & Tang, J. (2018) Network Embedding as Matrix Factoriza- tion: Unifying Deepwalk, LINE, PTE, and Node2Vec. In International Conference on Web Search and Data Mining.