BIG DATA ANALYSIS AND PROJECT
Assignment 1 – Part A

Milestone 1A

FORMULATION OF QUESTION AND PREPROCESSING

Kishaiyan Vellaichamy Thangaraj
a1819309

INTRODUCTION

In recent years, the intersection of mental health and economic productivity has garnered increasing attention from researchers, policymakers, and health professionals. Mental health problems have a significant impact on people's well-being as well as the global economy. These problems can range from anxiety and depression to more serious diseases like bipolar disorder and schizophrenia. The World Health Organization (WHO) notes that mental health issues have a substantial impact on people's capacity to conduct daily activities, including work, and are among the primary causes of disability globally.

The financial toll that mental health disorders take is enormous. The National Institute of Mental Health (NIMH) estimates that lost productivity from mental health disorders costs the world economy $1 trillion every year in the US. This loss shows up in many things, such as less productivity at work, higher absenteeism, and higher turnover rates. Workers with mental health disorders are more likely to perform poorly at work, which lowers an organization's overall production and efficiency.

Moreover, direct costs associated with mental health include more than just medical bills and treatment. The financial burden on countries is compounded by indirect costs, such as the loss of skilled workers and the long-term effects on workforce participation. The UN has also emphasized the significance of addressing mental health as a crucial element of accomplishing sustainable development goals, realizing that nations' capacity to maintain economic stability and progress depends on their ability to maintain good mental health.

Considering these difficulties, using big data to examine population patterns in mental health is a viable way to comprehend and lessen the effects of mental health on both a personal and a financial level. Big data analytics can help with more effective interventions and policies by revealing patterns and connections that traditional research approaches might miss. To inform measures to improve mental well-being and economic resilience, this project intends to use big data to investigate the prevalence, causes, and effects of mental health concerns.

Hypothesis:

The mental health of the people is affected by the heightened social media usage, economic instability, prolonged effects of the COVID-19 pandemic and work-life balance which in turn affects the economy

To examine the hypothesis concerning the rise in mental health problems among people, we will be utilizing a range of datasets that jointly demonstrate big data attributes. Volume, variety, velocity, and veracity—the four V's of big data—are the reasons behind this classification.

1. **Volume:**
    - Large volumes of data gathered over time from many sources, including social media platforms, economic indicators, health records, and job information, are included in the datasets that were selected. Millions of transactions and records are covered by economic data, whereas billions of interactions and postings are involved in social media usage data.
2. **Variety:**
    - Structured data (like unemployment rates and COVID-19 case numbers) and unstructured data (like social media posts and text-based mental health surveys) are the sources of the data. This diversity necessitates advanced techniques to combine and evaluate various data formats to produce thorough insights.
3. **Velocity:**
    - Particularly in the case of real-time social media interactions and continuous health monitoring data, the data is generated and updated at fast speeds. Additionally updated regularly to reflect ongoing environmental changes are statistics about pandemics and the economy.
4. **Veracity:**
    - Ensuring the data's dependability and accuracy is essential. Misinformation can be found in social media data, reporting errors can affect economic statistics, and there may be anomalies in health data. To solve these problems and preserve the analysis's integrity, procedures for data cleaning and validation must be put in place.

PRE-PROCESSING

The data collected for this analysis are from credible sources such as government and inter-governmental organizations. To ensure the data is suitable for answering the hypothesis, several pre-processing techniques have been applied:

Data Cleaning:

- Handling Missing Values: Removing records that have a sizable quantity of missing data and imputing missing values using the proper techniques, such as mean imputation for numerical data or mode imputation for categorical data.
- Removing Duplicates: To avoid bias in the analysis, duplicate records should be found and removed.

Data Transformation:

- **Normalization and Standardization:** Data must be standardized to guarantee that it follows a standard distribution and to bring numerical data into accordance with a common scale without distorting variations in value ranges.
- **Encoding Categorical Variables:** Converting categorical data into numerical formats using techniques such as one-hot encoding or label encoding, which are essential for machine learning algorithms.

Data Integration:

- **Merging Data Sources:** creating a comprehensive dataset by combining datasets from several sources based on shared keys or attributes, guaranteeing consistency throughout all integrated data.
- **Addressing Data Inconsistencies:** Resolving conflicts and discrepancies across sources' definitions, data formats, and measurement units.

Applying these pre-processing techniques ensures that the data is clean, consistent, and ready for in-depth analysis. These steps are crucial for building accurate and reliable models to test the hypothesis and derive meaningful insights into the factors affecting the recent spike in mental health issues.

REFERENCES

- World Health Organization (WHO). (n.d.). Mental health in the workplace. Retrieved from https://www.who.int/news-room/fact-sheets/detail/mental-health-in-the-workplace
- National Institute of Mental Health (NIMH). (n.d.). Mental Health Information. Retrieved from https://www.nimh.nih.gov/health/statistics/mental-illness
- Harvard Business Review. (2021). The Hard Facts About Mental Health in the Workplace. Retrieved from https://hbr.org/2021/10/the-hard-facts-about-mental-health-in-the-workplace
- United Nations. (2020). Mental Health Matters: Social Inclusion of Youth with Mental Health Conditions. Retrieved from https://www.un.org/development/desa/youth/news/2020/02/mental-health/