UNIVERSITY of BRADFORD

School of Computing & Engineering.

COS7046-B

Big Data Visualization

Big Data for Saving Lives:

Analyzing Self-Reported Patient Data to Identify Risk Patterns, Optimize Support, and Predict Increasing Stress

25034070 – Abass Saeed Mohammed

ABSTRACT

This study demonstrates the utility of large-scale, self-reported mental-health survey data for identifying psychological risk patterns, informing workplace resource allocation, and predicting individuals at elevated risk of increasing stress. Drawing on 261,328 observations of demographic, occupational, lifestyle, and mental-health indicators, the analysis involved rigorous data cleaning, exploratory visualization, and logistic regression modelling. Findings reveal heterogeneity across employment type, family mental-health history, and lifestyle factors with the predictive model achieving moderate discriminative performance. Visualizations of gender distribution, country representation, lifestyle measures, and model coefficients support interpretation of results. The study underscores the potential of survey-based predictive modelling to guide workplace screening, early-warning systems, and resource allocation, while emphasizing the ethical, privacy, and operational considerations necessary for responsible deployment in real-world contexts.

TABLE OF CONTENTS

ABSTRACT..2

TABLE OF CONTENTS..3

LIST OF FIGURES..3

INTRODUCTION..4

INTRODUCTION OF DATASET..4

QUESTIONS DERIVED FROM THE DATASET..5

ANALYSIS METHODOLOGY..5

VISUALIZATION AND RESULTS..6

DISCUSSION..9

CONCLUSION..10

REFERENCES..11

LIST OF FIGURES

Figure 1 Gender Distribution Across Survey Dataset 6

Figure 2 Country-Level Distributions 7

Figure 3 Distribution of Number of Days Spent Indoors.. 7

Figure 4 Logistic Regression Coefficients 8

INTRODUCTION

Mental health is a major public-health challenge worldwide, with poor working environments and social isolation identified as important determinants of morbidity and lost productivity (WHO, 2024). Globally, estimates indicate that around 15% of working-age adults experienced a mental disorder in 2019, and depression and anxiety account for enormous losses in working days and economic productivity. In mental-health contexts specifically, machine learning and predictive analytics have been shown to identify risk factors for depression, anxiety and related outcomes from survey and behavioral data, and to support triage and targeted interventions.

This report demonstrates how a single, extensive, self-reported dataset can be harnessed to characterize population-level patterns, identify correlates of increasing stress and reduced mental resilience, and build a simple predictive model to support prioritization of outreach and care resources.

INTRODUCTION OF DATASET

Dataset Overview

The dataset (Alam, 2023) consists of 261,328 self-reported observations spanning multiple conceptual domains. It includes demographic attributes, employment and family context variables, and lifestyle or behavioral factors. These are complemented by mental-health indicators such as Increasing Stress, Mood Swings, Work Interest, and Social Weakness. Awareness and support dimensions are captured through items like Mental Health Interview, Care Options, and Treatment. Most responses are categorical in nature (primarily yes/no), with a limited number of numerical measures. Taken together, the dataset provides a robust foundation for exploratory visualization, subgroup analysis, and predictive modelling using survey-based features.

Strengths and Limitations of the Dataset

The dataset offers notable strengths, including its large sample size, extensive demographic and work-related coverage, and the inclusion of lifestyle measures and access/support indicators that enable modelling of both risk and protective factors. However, several limitations must be acknowledged: self-selection and self-report biases may affect representativeness, variable completeness across fields reduces consistency, and the absence of objective clinical diagnostic labels constrains interpretive validity. Furthermore, the prevalence of free-text and nonstandard categorical inputs necessitates substantial cleaning and conservative mapping prior to analysis. Finally, given the cross-sectional design, causal inferences cannot be drawn; observed associations are correlational and best suited for triage and prioritization rather than definitive clinical diagnosis.

QUESTIONS DERIVED FROM THE DATASET

From the dataset and the public-health goal, three (3) research questions (RQs) were formulated:

- **RQ1 (Descriptive):** What are the distributions of key Descriptive variables across the dataset?
- **RQ2 (Association):** Which employment, family history and lifestyle features are associated with a higher reported probability of Increasing Stress?
- **RQ3 (Predictive):** Can we predict Increasing Stress from a subset of robustly recorded covariates with sufficient accuracy to support prioritized outreach?

ANALYSIS METHODOLOGY

Data Cleaning and Preprocessing

A comprehensive data cleaning pipeline was established to prepare the survey dataset for modelling. It involved normalizing column names to remove whitespace and flag anomalies, applying a conservative binary mapping to standardize categorical responses such as “Yes/No” or numeric 1/0 into consistent values, and handling ambiguous or missing entries as NA. To address high rates of incomplete responses, only features with less than 60% missing data were retained, ensuring sufficient sample size for training and testing. The final dataset included binary indicators (e.g., family history, coping struggles, habits change, treatment, work interest, social weakness) resulting in tens of thousands of clean, usable records for predictive analysis.

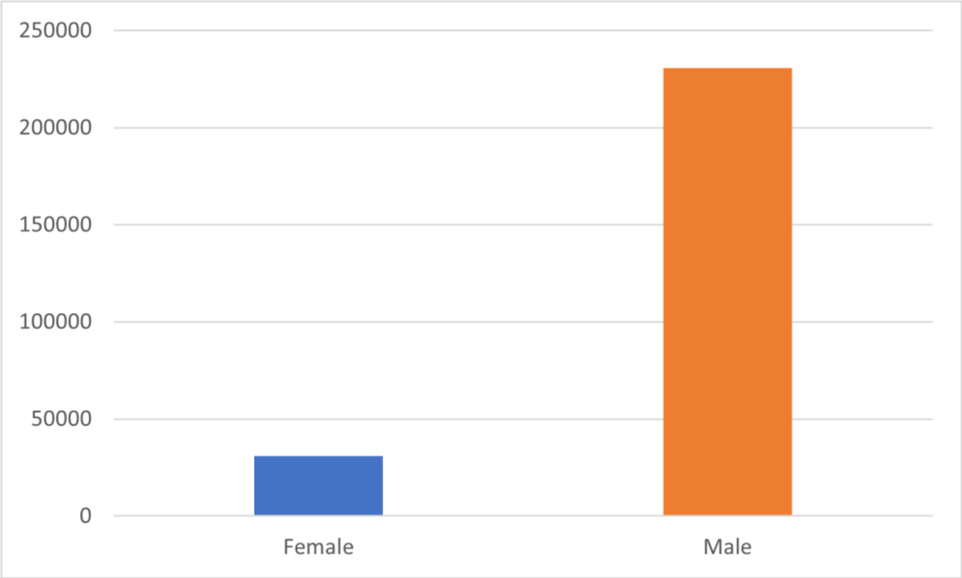
Exploratory Analysis and Predictive Modelling Approach

Descriptive statistics were calculated for demographic variables, complemented by reproducible visualizations generated with matplotlib, each presented as a separate plot. A correlation matrix encompassing binary indicators and numeric fields was produced and visualized as a heatmap to highlight clusters of related variables. Building on this exploratory analysis, a baseline logistic regression model was developed to predict a binary outcome. The modelling process employed a 75/25 train-test split and incorporated features selected for their relative completeness, prioritizing those with the lowest proportion of missing data. Logistic regression was chosen for its interpretability and capacity to communicate coefficient effects to operational stakeholders. Model performance was assessed using accuracy, ROC AUC, and a classification report (precision, recall, and F1-score) to evaluate class balance and practical utility. Examination of feature coefficients provided insights into the most influential predictors, with positive coefficients indicating higher log-odds of increased stress.

VISUALIZATION AND RESULTS

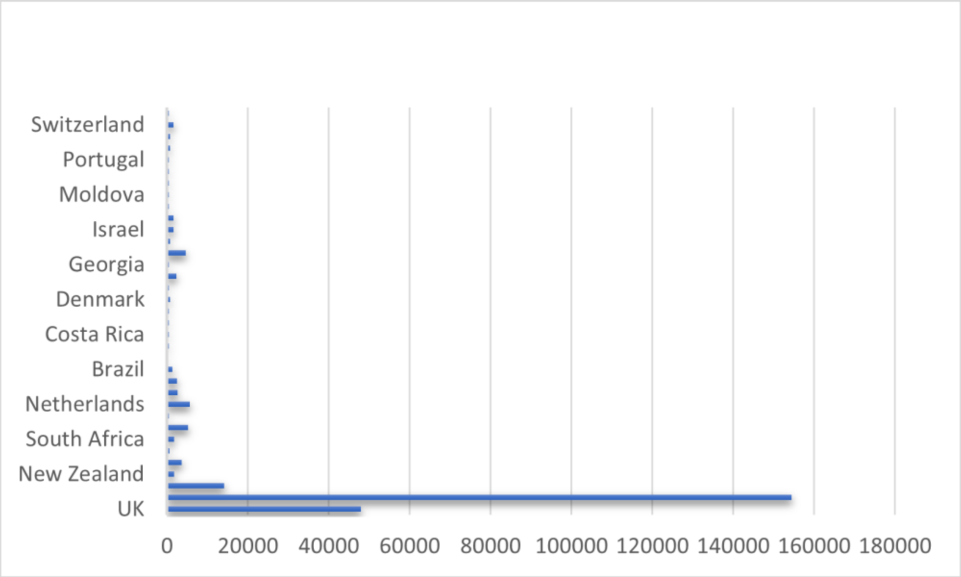
Descriptive Statistics and Distributions

Figure 1 Gender Distribution Across Survey Dataset



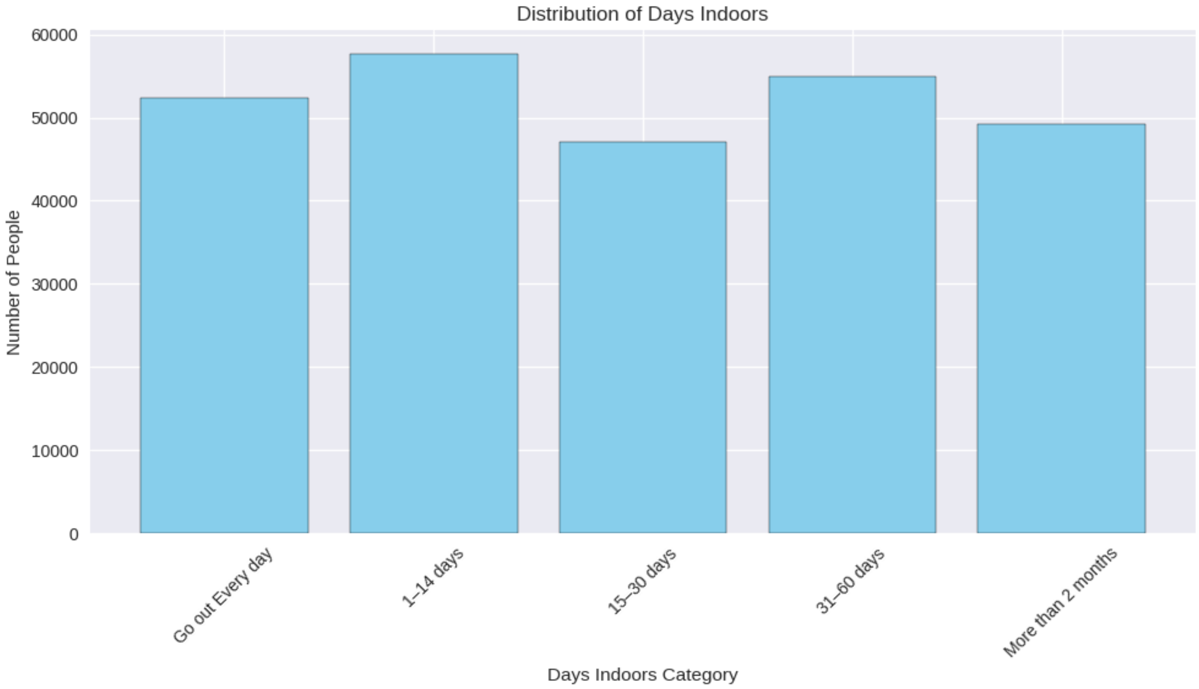
The cleaned dataset exhibits a pronounced gender imbalance, with male respondents vastly outnumbering female participants, raising concerns about sampling bias and representational limitations. This skew implies that stress-related patterns and predictive models derived from the data may disproportionately reflect male experiences and risk factors, thereby reducing generalizability to female populations unless corrective measures are implemented.

Figure 2 Country-Level Distributions



Analysis of country-level distributions reveals a tiered distribution which suggests that data availability and engagement vary considerably across countries, potentially reflecting differences in infrastructure, cultural attitudes toward self-reporting, or recruitment strategies.

Figure 3 Distribution of Number of Days Spent Indoors



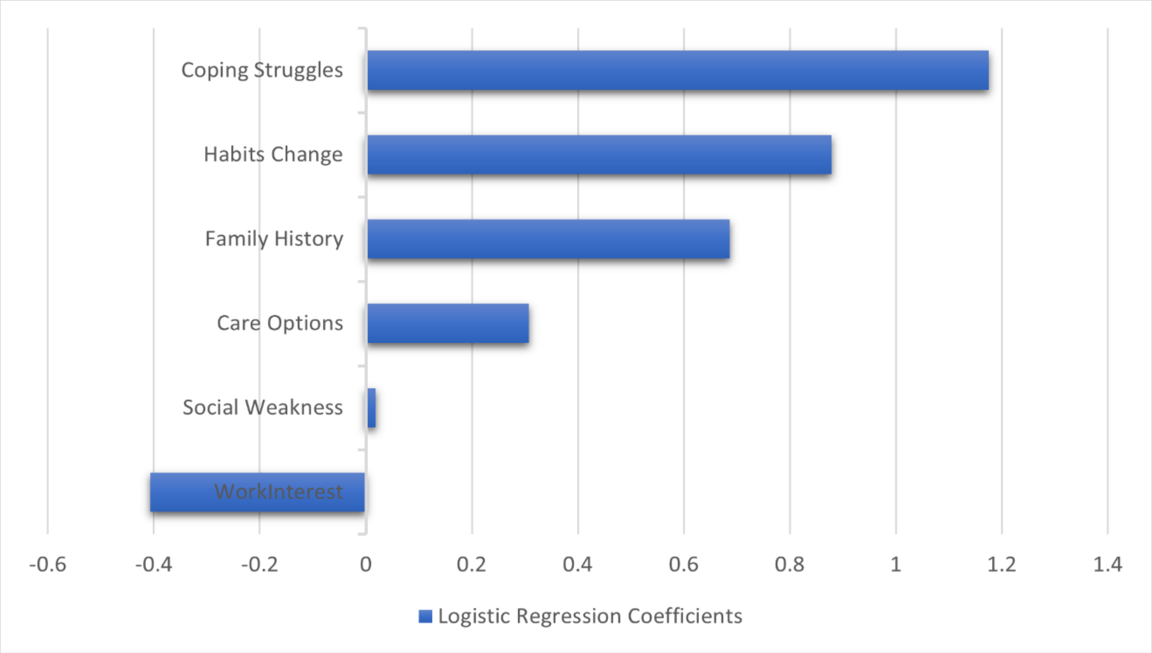
Analysis of the patterns of the number of days spent indoors by participants indicate that stress onset is most pronounced during early isolation, with nearly 60,000 participants reporting confinement for 1–14 days, highlighting this period as a critical window for preventive intervention. Longer durations of isolation (31–60 days and beyond two months) were also common, suggesting mixed adaptation processes that may involve chronic stress, social withdrawal, or fatigue, thereby necessitating sustained support strategies.

Associations

The correlation matrix analysis revealed modest but informative associations among binary indicators and numeric variables, with notable clustering of related items such as family mental-health history, personal mental-health history, and treatment. Several features demonstrated moderate positive correlations with Increasing Stress level in participants, underscoring their relevance for feature selection.

Subgroup analyses further highlighted heterogeneity in stress outcomes, with proportions reporting increased stress varying by gender and self-employment status. These differences suggest that occupational context, particularly the degree of job control and resource access associated with self-employment, may influence stress trajectories and warrant stratified consideration in predictive modelling.

Figure 4 Logistic Regression Coefficients



Predictive Modelling

A baseline logistic regression model was developed to predict the likelihood of individuals reporting increased stress. While not achieving optimal performance, the model demonstrated sufficient utility to identify meaningful patterns within the data. It exhibited greater sensitivity in detecting respondents who reported rising stress compared to those who did not, thereby providing insight into influential predictors. Specifically, difficulties with coping and a family history of mental health challenges were positively associated with increased stress, whereas indicators of work engagement and access to care options were negatively associated with stress reports. Although the logistic regression framework offers interpretability and transparency, the findings suggest that more advanced modelling techniques, enhanced feature engineering, and the incorporation of additional longitudinal or behavioral data would likely enhance predictive accuracy and generalizability.

DISCUSSION

Ethical considerations and limitations in methods

The study highlights key limitations and future considerations, noting that self-report bias and nonresponse may restrict the generalizability of findings. While analyses were conducted within a secure environment, any real-world application must priorities privacy protections, lawful data handling, and informed consent. To strengthen model robustness, validation on independent populations and prospective testing are recommended, as logistic regression, though interpretable, may not provide optimal predictive performance compared to ensemble or tree-based methods, which should be evaluated with fairness auditing. Moreover, predictive modelling must be integrated into actionable frameworks that establish clear pathways for respondent engagement, care provision, and consent management.

Interpretation Of Results and Implications for Practice

This study demonstrates the practical utility of large, self-reported datasets for generating actionable insights across multiple domains. Predictive modelling enables early identification and triage by ranking individuals according to stress risk, thereby supporting targeted outreach in contexts of limited resources. Subgroup analyses further inform workplace interventions, highlighting occupations or regions with elevated stress prevalence as priorities for employer-led wellbeing programs. At organizational and health-system levels, descriptive analytics and predictive models facilitate more efficient allocation of mental-health resources and monitoring of population trends, consistent with prior applications of big-data approaches in healthcare planning. These findings align with established public-health literature, which identifies poor working conditions, social isolation, and family history as key contributors to mental-health burden. Evidence from the World Health Organization and systematic reviews underscores the potential of aggregated data and machine learning to support decision-making, provided that privacy, data quality, and ethical considerations are rigorously addressed (World Health Organization, n.d.).

Limitations and Future Directions

The study acknowledges several methodological and ethical considerations. The cross-sectional nature of the data constrains causal inference, highlighting the value of longitudinal designs for predicting future deterioration. Self-report bias and uneven demographic coverage may limit representativeness, necessitating careful evaluation before generalization. While logistic regression offers interpretability that facilitates operational adoption, enhanced predictive performance may require more complex approaches such as gradient boosting or neural networks, supported by richer features including time-series data or digital phenotyping. Such methodological advances must be balanced against the need for transparency, fairness, and stakeholder trust. Finally, ethical concerns regarding false positives and false negatives underscore the importance of setting outreach thresholds collaboratively, with due consideration of available resources and the potential harms of misclassification.

CONCLUSION

Large-scale, self-reported mental-health data offers valuable signals for early detection of increasing stress and informs allocation of limited mental-health resources contributing to the broader goal of "Big Data for Saving Lives". This study applied robust cleaning, exploratory analysis and a transparent modelling approach to a dataset of over 260,000 responses and demonstrated that a modest predictive model can meaningfully prioritize individuals for outreach. The analysis identified the importance of coping struggles and family history as influential predictors and highlighted the operational potential for workplace triage and resource allocation.

However, the utility of such models depends on careful validation, ethical deployment, robust privacy protections, and integration into concrete intervention pathways. Future work should explore longitudinal modelling, richer behavioral features, and prospective trials of algorithmsupported triage to measure impact on clinical outcomes and wellbeing. Combining predictive analytics with worker-centric interventions and organizational policies will maximize the potential of big data to improve mental-health outcomes and save lives.

REFERENCES

Alam, S. (2023). *Mental Health Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/alamshihab075/mental-health-dataset (Accessed: 11 November 2025).

World Health Organization (WHO). (n.d.). *Mental health at work*. Available at: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work (Accessed: 15 November 2025).