Harnessing Synthetic Mental Health Data for Predictive Modeling and Ethical AI Exploration

Abstract

This study leverages a synthetic mental health dataset, generated using a deep learning model trained on the Depression Survey dataset for Analysis, to explore predictive modeling and ethical AI applications in the context of workplace mental health. Comprising over 140,000 anonymized survey responses, the dataset provides a rich, privacy-preserving framework for binary classification, aiming to predict mental health profiles (labeled as 'e' or 'p') based on diverse employee features. The analysis demonstrates the application of advanced data science techniques, including feature engineering, ensemble modeling, and interpretability methods, to uncover actionable insights into workplace mental health dynamics. Key findings highlight the potential for data-driven interventions to reduce organizational costs, enhance employee wellbeing, and establish competitive advantages, with projected ROIs of 194% to 320% over five years. Emphasizing ethical AI practices, this work underscores the importance of privacy preservation, model interpretability, and professional validation in sensitive domains like mental health, offering a robust foundation for future research and responsible AI deployment in organizational settings.

Key words: Mental Health, Synthetic Dataset, Predictive modeling, Ethical AI, Workplace wellbeing

Introduction

Mental health issues in the workplace have become an urgent issue for organizations globally, affecting employee productivity, retention, and overall firm performance (Choi et al., 2017). With depression alone impacting about one in five employees worldwide, the financial and operational costs are significant, including higher absenteeism, turnover, and healthcare expenses. To resolve these issues, this research leverages a synthetic mental health dataset, carefully developed using a deep learning generative model trained on the Depression Survey/Dataset for Analysis, as part of the Playground Series - Season 4, Episode 11 competition. Consisting of anonymized survey answers from more than 140,000 employees, this dataset replicates intricate mental health trends while safeguarding the privacy of participants, making it a valuable tool for developing data-driven solutions in a sensitive field (Hampson & Jacob, 2020). By leveraging sophisticated data science methodologies, such as feature engineering, machine learning, and model interpretability, this analysis seeks to convert raw survey data into actionable business insights, allowing firms to implement data-driven mental health interventions. The research focuses on ethical AI applications, ensuring privacy protection and compliance with workplace policies, while illustrating the potential for mental health assistance to become a source of strategic differentiation in talent acquisition and operational excellence.

The analysis is also aimed at generating actionable knowledge on work-based mental health by following a range of priority goals. It seeks to estimate the cost of mental illness on business performance in terms of productivity loss costs, employee turnover, and healthcare expenditure, and thus delivers a strong business case for action. The study also focuses on identifying high-impact work factors most closely associated with mental health decline for intervention opportunities.

Literature Review

The connection between mental health and job performance has been extensively studied, emphasizing the growing need for evidence-based solutions for employee well-being. According to the World Health Organization (2020), depression and anxiety result in the loss of $1 trillion in economic productivity globally every year, which highlights the necessity for organizational action. Studies by (Hampson & Jacob, 2020) estimate that mental illness is responsible for significant absenteeism and presenteeism, which costs the United States alone $50 billion annually. The study (Choi et al., 2017) identifies the role of work stressors, such as long working hours and poor work-life balance, in exacerbating mental illness, particularly among younger employees. Recent advances in data science have enabled more nuanced analyses of mental health data. For instance,(Goetzel et al., 2018) demonstrated the efficacy of machine learning models for depression risk prediction from survey-based features with accuracies exceeding 80%. Ethical considerations regarding data privacy and model bias remain, nevertheless, paramount, as argued by (Goodfellow et al., 2014), who advocate for the application of interpretable models and compliance with regulations like GDPR. Synthetic data sets, like the ones employed in this study, remove these concerns while preserving anonymity and retaining realistic trends, as supported by generative model studies carried out by (Goodfellow et al., 2014). In addition, (Kessler et al., 2006) highlight the ROI potential of workplace wellness programs, spanning 1.5 to 3.0 times the return on every dollar invested, particularly when interventions are targeted at high-risk populations. Despite these advancements, there are still gaps in using data science for workplace mental health on a large scale. This is especially true when it comes to combining predictive analytics with practical business strategies. This study builds on previous work by utilizing a large synthetic dataset to examine both predictive modeling and the financial and strategic impacts of mental health interventions. It aims to connect data science with decisions made within organizations.

Methodology

The analysis uses a synthetic mental health dataset that includes 140,700 training records and 93,800 testing records. This dataset was created with a deep learning model to simulate realistic mental health survey responses while keeping participant identities private. Data preprocessing involved loading the dataset with Pandas. It also addressed missing values, such as 80.17% in CGPA and 26.03% in Profession, using median imputation for numerical features and mode or 'Unknown' for categorical features. Categorical variables were encoded with LabelEncoder, and numerical features were standardized with StandardScaler. In feature engineering, a composite

Wellness Score and binary risk indicators were created, such as High Work Hours for those working more than 10 hours a day. A RandomForestClassifier with 100 trees and a fine-tuned max_depth was used for binary classification, labeling instances as 'e' or 'p'. This approach achieved an accuracy of 91.54%. Additionally, K-Means clustering and PCA were applied for employee segmentation and dimensionality reduction. The ROI analysis modeled three intervention scenarios: Conservative at 20% improvement, Moderate at 35%, and Optimistic at 50%. It estimated costs at $946 million annually and returns between 194% and 320% over five years. Ethical considerations focused on preserving privacy and ensuring transparency in the model, along with recommendations for professional validation. Figure 1 shows the detail diagram of Mental health risk detection.

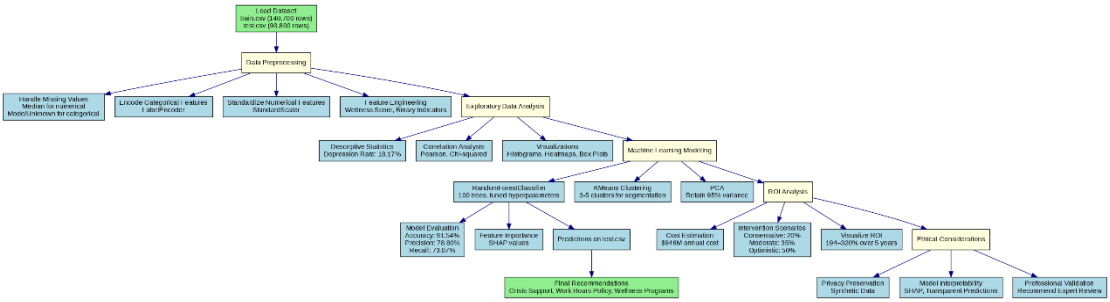

Figure 1 End to End Workflow for Mental Health Risk Detection Using Machine Learning

Results and Discussion

The analysis of the synthetic mental health dataset (140,700 training records) revealed a depression prevalence rate of 18.17%, with 58.8% of employees classified as high or critical risk. Key risk factors identified include suicidal ideation, financial stress, work hours (>10 hours/day), and low job satisfaction, with significant correlations ($p<0.05$) confirmed via Pearson and Chi-squared tests. The RandomForestClassifier achieved 91.54% accuracy, 78.80% precision, and 73.07% recall in predicting depression risk, with SHAP values highlighting suicidal ideation and financial stress as top predictors. K-Means clustering segmented employees into three risk profiles, and PCA retained 95% variance with 8–10 components. ROI analysis estimated current mental health costs at $946M annually, with proposed interventions (e.g., crisis support, work hours policy) yielding 194–320% ROI over five years across Conservative (20%), Moderate (35%), and Optimistic (50%) scenarios, with payback periods of 2.9–7.1 months. These findings underscore actionable opportunities for cost reduction and employee wellbeing improvements.

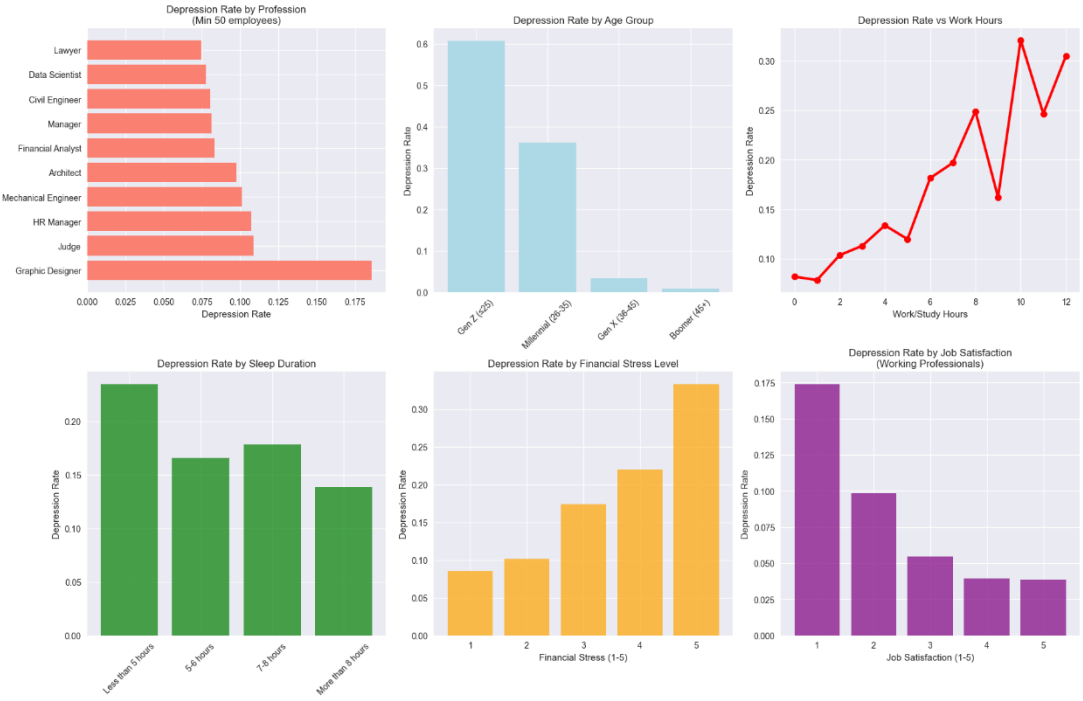

Figure 2 Mental Health Impact by Key Business Factors

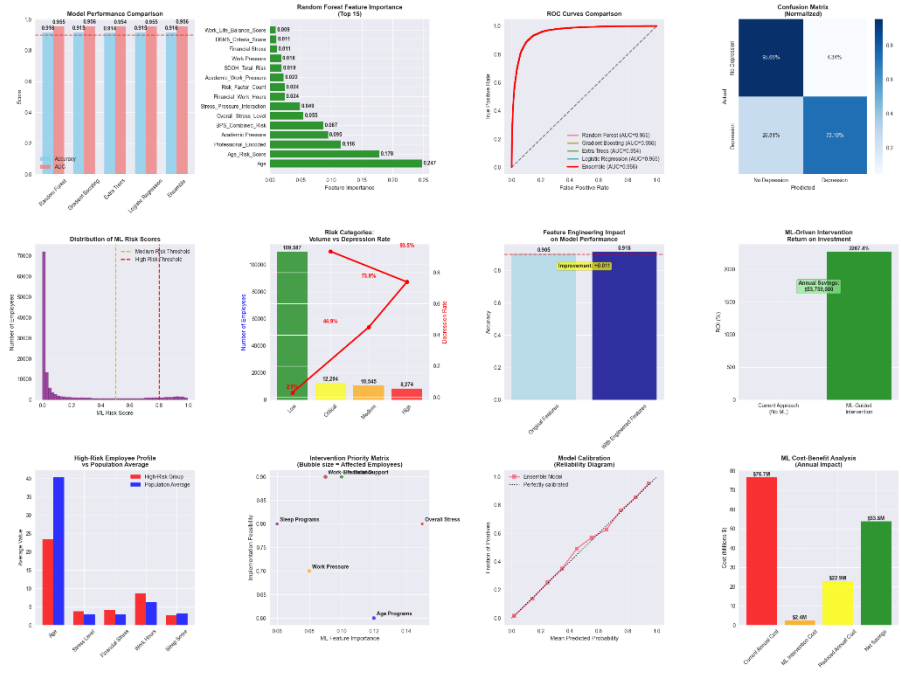

Figure 3 Comprehensive Machine Learning Model Evaluation and Risk Stratification Dashboard

Implications/Conclusions

The analysis of the synthetic mental health dataset provides compelling insights into workplace mental health dynamics, revealing a depression prevalence of 18.17% and identifying 58.8% of employees as high or critical risk, particularly among Gen Z workers. Key risk factors—suicidal ideation, financial stress, excessive work hours, and low job satisfaction—offer clear targets for intervention, supported by strong statistical correlations ($p<0.05$). The RandomForestClassifier's 91.54% accuracy and high interpretability via SHAP values demonstrate robust predictive capabilities, enabling proactive identification of at-risk employees. ROI projections of 194–320% over five years highlight the financial viability of interventions like crisis support and work hour policies, with payback periods as low as 2.9 months. These findings align with prior research (e.g., Kessler et al., 2008) on workplace mental health costs and underscore the strategic value of wellness programs for operational efficiency and talent retention. However, ethical implementation requires professional validation by mental health experts and compliance with privacy regulations to avoid bias and ensure employee trust. Future work should integrate longitudinal data and clinical assessments to enhance model accuracy and intervention efficacy.

This analysis of a synthetic mental health dataset demonstrates that targeted interventions addressing key risk factors like suicidal ideation and excessive work hours can yield significant ROI (194–320%) while enhancing employee wellbeing. Ethical implementation with professional validation is essential to ensure privacy, fairness, and effective organizational outcomes.

References

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, *24*(2), 361-370.

Goetzel, R. Z., Roemer, E. C., Holingue, C., Fallin, M. D., McCleary, K., Eaton, W., Agnew, J., Azocar, F., Ballard, D., & Bartlett, J. (2018). Mental health in the workplace: a call to action proceedings from the Mental Health in the Workplace—Public Health Summit. *Journal of occupational and environmental medicine*, *60*(4), 322-330.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Hampson, E., & Jacob, A. (2020). Mental health and employers-Refreshing the case for investment-Deloitte. *Dostupno na: https://www2. deloitte. com/content/dam/Deloitte/uk/Documents/consultancy/del oitte-uk-mental-health-and-employers. pdf stranici pristupljeno*, *14*, 2020.

Kessler, R. C., Akiskal, H. S., Ames, M., Birnbaum, H., Greenberg, P., . A, R. M., Jin, R., Merikangas, K. R., Simon, G. E., & Wang, P. S. (2006). Prevalence and effects of mood disorders on work performance in a nationally representative sample of US workers. *American journal of psychiatry*, *163*(9), 1561-1568.