International Journal of Science and Mathematics Education

E-ISSN: 3062-7214
P-ISSN: 3062-7206

Review Article

Optimizing IT Remote Workers Mental Health Prediction using Feature Engineering

Fikri Muhamad Fahmi [1], Budiman [2*], Nur Alamsyah [3]

[1] Universitas Informatika dan Bisnis Indonesia, Indonesia 1;
[2] Universitas Informatika dan Bisnis Indonesia, Indonesia 2; e-mail : budiman@unibi.ac.id
[3] Universitas Informatika dan Bisnis Indonesia, Indonesia 3;

* Corresponding Author : Fikri Muhamad Fahmi

Abstract: Given the increasing prevalence of mental health challenges in digital work settings, especially among IT remote workers, early detection mechanisms have become critically important. This study aims to improve the prediction accuracy of mental health conditions among IT remote workers by integrating feature engineering techniques within machine learning models. Five algorithms consisting of Random Forest, Logistic Regression, K-Nearest Neighbors, Decision Tree, and Naive Bayes were evaluated. The Random Forest model achieved the best performance, with 83% accuracy, 83% precision, 100% recall, and a 90% F1-score, followed closely by Logistic Regression with 82% accuracy. Nevertheless, the results demonstrate the feasibility of applying machine learning to support the early detection of mental health risks, offering a strong foundation for future research in predictive analytics and the development of intelligent support systems within digital work environments.

Keywords: Random Forest, Classification, Machine Learning, Mental Health

1. Introduction

Remote working has been one of the major changes in the modern world of work, especially since 2019. Remote working is a work system that allows employees to work remotely, and not tied to a specific location. Previously, this work system was only implemented by a few organizations or companies with the technological infrastructure and human resources to support it. According to (McKinsey, 2021), the COVID-19 pandemic has driven the trend of remote working 25% higher than previously estimated, and many organizations are continuing this policy after the pandemic because it is considered effective, efficient, and offers greater flexibility. This transition has been associated with both positive and negative effects on workers' well-being, necessitating a deeper understanding of these dynamics to develop effective strategies for mental health support. Remote work has been linked to increased mental health issues, including anxiety, depression, and sleep disorders. Studies have shown that individuals facing difficulties with remote work report higher rates of these conditions compared to those who do not experience such challenges (Nowrouzi-Kia et al., 2024; Sim et al., 2024). The lack of social interaction and professional isolation are significant contributors to these mental health challenges (Charalampous et al., 2018; Lyzwinski, 2024). Several factors influence the mental health of remote workers. Motivation plays a crucial role, with higher motivation levels associated with lower psychological stress

Received: March, 13th 2025
Revised: April, 01th 2025
Accepted: April, 23th 2025
Published: June, 2025

Curr. Ver.: June, 2025

(Idaiani & Waris, 2022). Additionally, the ability to maintain a work-life balance is critical, as remote work can blur the lines between professional and personal life, leading to increased stress and decreased well-being (Nowrouzi-Kia et al., 2024; Sarinastiti et al., 2022). Machine learning has been extensively used for the detection and diagnosis of mental health conditions such as depression, anxiety, and schizophrenia. Techniques like support vector machines, decision trees, and neural networks are commonly employed to analyze various data types, including medical records and social media interactions, to identify symptoms and classify the severity of mental health conditions (Glaz et al., 2019; Nash et al., 2023; Shatte et al., 2019).

This study aims to enhance prediction accuracy for IT remote workers' mental health conditions by integrating feature engineering within machine learning models. Feature engineering uncovers key factors contributing to mental health disorders by refining input variables, ensuring reliable performance on new data. This approach significantly enhances model accuracy for better decision-making. Additionally, it provides a solid foundation for future research, advancing predictive analytics and supporting business growth in a dynamic digital landscape.

2. Literature Review

According to (Sofianti et al., 2023), the transition from conventional work systems to remote work presents challenges, including adapting to new problems and also balancing personal life and work life or work-life balance, as well as the mental health of individuals. Recent studies have explored machine learning approaches for predicting mental health disorders, particularly in remote and technical workplaces. Various classification models, including Decision Trees, Random Forest, and Logistic Regression, have been employed with hyperparameter tuning to optimize performance (Kaushik et al., 2024; Mohammad & Siddiqui, 2021) Feature selection techniques like Recursive Feature Elimination and LASSO have been utilized to identify the most impactful factors contributing to mental health outcomes (Mallick & Panda, 2024). These models have achieved accuracies ranging from 82% to 95% using different optimization techniques such as Grid Search CV, Bayesian Optimization, and Optuna (Cheng & Haw, 2023; Mohammad & Siddiqui, 2021). The Random Forest model, in particular, has shown promising results with accuracies of 83-89% (Cheng & Haw, 2023; Kaushik et al., 2024). These AI-driven approaches can complement traditional diagnostic methods, potentially improving early detection and intervention strategies for mental health disorders in workplace settings (Kaushik et al., 2024). While existing studies have demonstrated the effectiveness of various machine learning models in predicting mental health outcomes, a structured methodology is essential to ensure a systematic and reproducible approach to data analysis.

3. Research Methods

Figure 1 illustrates the flow of research methods applied in this study to create a prediction model in remote workers' mental health condition using random forest. This flow consists of a series of interrelated steps, from data collection to select best model.

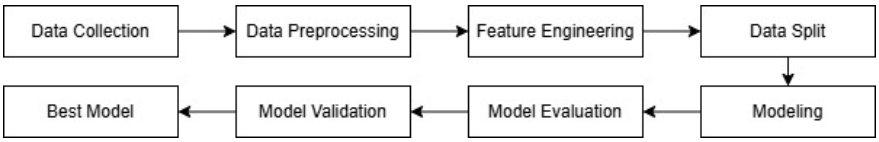

Figure 1. Research Method

3.1 Data Collection

In this study, we used the publicly available dataset from kaggle, which consist of several features related to the remote workers' mental health condition. The dataset has 5000 records and 20 features related to risk factors and other key contributors that can affect remote workers' mental health condition. Table 1 is a description of the remote workers' dataset.

Table 1. Data Description

Feature	Description
Employee_ID	Workers' identification or metadata.
Age	The age of the respondent.
Gender	The gender of the respondent.
Job_Role	The job role of the respondent.
Industry	The industry of the respondent.
Years_Of_Experience	Years of experience of the respondent.
Work_Location	Respondent work location consisting (remote/onsite/hybrid)
Hours_Worked_Per_Week	How much hour the respondent work in a week.
Number_Of_Virtual_Meetings	Number of virtual meeting attended in a week.
Work_Life_Balance_Rating	Work-life balance rating of the respondent.
Stress_Level	Respondent stress level
Mental_Health_Condition	Respondent mental health history (none/burnout/anxiety/depressed)
Access_To_Mental_Health_Resources	Access to a mental health facility (yes/no)
Productivity_Change	Productivity change after remote work.
Social_Isolation_Rating	Social isolation rating after remote work.
Satisfaction_With_Remote_Work	Respondent satisfaction of remote work.
Company_Support_for_Remote_Work	Company support rating for remote worker
Physical_Activity	Respondent physical activity frequency (none/weekly/daily).
Sleep_Quality	Respondent daily sleep quality (poor/moderate/good).
Region	Respondent current region (North America/Europe/Africa/Asia/Oceania).

3.2 Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for model development. This process involves cleaning the data, handling missing values, encoding categorical variables, scaling numerical features, and removing any anomalies or outliers. Additionally, data transformations and normalization are applied where necessary to ensure the data is in a suitable format for analysis.

3.3 Feature Engineering

Feature engineering involves creating new features or modifying existing ones to enhance the predictive power of the model. This can include polynomial features, interaction terms, and domain-specific transformations. Feature selection techniques, such as correlation analysis and feature importance evaluation, are also applied to retain only the most relevant variables.

3.4 Data Split

The dataset is divided into training and testing subsets. Typically, a common split is 80% for training and 20% for testing. The training set is used to build and optimize the model, while the test set provides an unbiased evaluation of the model's performance. This approach ensures that the model's generalization ability is effectively assessed.

3.5 Modeling

Various machine learning algorithms are applied to the training data to create predictive models. Algorithms such as linear regression, decision trees, naïve bayes, knn classifier, and random forest are explored. Performance metrics like accuracy, precision, recall, and f1-score are used to assess the models.

3.6 Best Model

The best-performing model is identified by comparing the results of different algorithms. Factors such as model complexity, training time, and generalization ability are considered.

4. Results

The dataset underwent preprocessing to handle missing values, outliers, and inconsistencies. Since there's a "None" value at the Mental_Health_Condition and Physical_Activity features and treated as null by pandas. To handle that, the "None" at the Mental_Health_Condition then replaced with "Healthy", while the "None" at the Physical_Activity replaced with "Sedentary". Filtering was also applied to the dataset so that the dataset only contained remote worker respondents. All the numerical features then normalized using MinMaxScaller, as for the categorical features are transformed using LabelEncoder and OneHotEncoding.

Table 1. Conducted Feature Engineering

New Feature	Description
Age_Group	Age category based on the specified range (e.g., "30s" for ages between 30 and 40).
Experience_Level	Years of experience level based on the specified range (e.g., "Junior" for experience between 0-2 years).
Virtual_Meeting_Frequency	Meeting frequency based on the specified range (e.g., "Moderate" for meeting frequency between 4-12 time per week)
Hours_Worked_Per_Day	Hours_Worked_Per_Week divided by 5, assuming there's 5 work days.
Hours_Worked_Per_Experience	Hours_Worked_Per_Week divided by the Years_Of_Experience
Has_Mental_Health_Condition	Binary target feature creation based on the Mental_Health_Condition feature (if "Healthy" then 0, else then 1).

Table 2 shows the feature engineering conducted on this dataset. A total of 6 new features is created by transforming and combining several original features to improve model performance on predicting remote workers mental health condition. The dataset then splitted using 70:30 ratio, 70% for training set and 30% for test set.

Table 2. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	83%	83%	100%	90%
Logistic Regression	82%	84%	97%	90%
KNN	80%	83%	97%	89%
Decision Tree	72%	84%	82%	83%
Naive Bayes	68%	84%	77%	80%

Table 3 presents a comparative analysis of five machine learning models consisting of Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes—based on four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. Random Forest achieved the highest Recall (100%), making it the most effective at correctly identifying all relevant instances. It also had the highest overall Accuracy (83%) and a strong F1-Score (90%), indicating a good balance between precision and recall. Logistic Regression showed comparable performance to Random Forest, with slightly lower Recall (97%) but higher Precision (84%), resulting in the same F1-Score (90%). This suggests it is slightly better at minimizing false positives. KNN had a slightly lower Accuracy (80%) and F1-Score (89%), but still maintained strong Recall (97%), similar to Logistic Regression. Its lower precision

(83%) slightly reduced its overall F1-Score. Decision Tree showed moderate performance with Accuracy (72%) and a balanced Precision (84%) and Recall (82%), resulting in an F1-Score (83%). Naive Bayes performed the weakest among the models, with the lowest Accuracy (68%), Recall (77%), and F1-Score (80%), despite having a decent Precision (84%). In summary, Random Forest and Logistic Regression outperformed the other models in terms of overall effectiveness, with strong F1-Scores and Recall, making them the most reliable for this classification task.

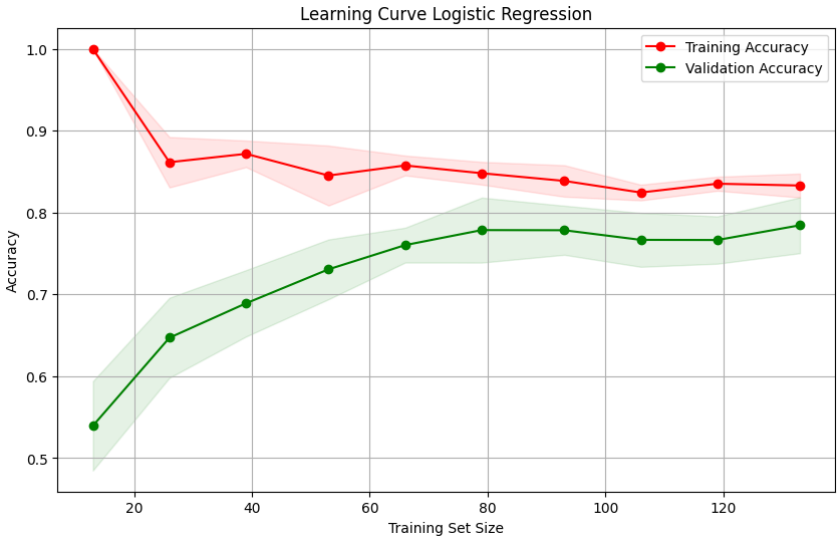

Figure 2. Learning Curve for Logistic Regression

Figure 2 illustrates how the Logistic Regression model improves as the training dataset increases. Initially, the training accuracy (red line) is high with small datasets, indicating that the model memorizes the data, which often suggests a tendency toward overfitting. As the volume of data grows, training accuracy slightly decreases, reflecting the model's transition toward learning more generalizable patterns and reducing the risk of overfitting.The validation accuracy (green line), which begins lower than the training score, gradually increases with additional data. This trend indicates that the model becomes better at generalizing to unseen data. Eventually, the training and validation scores converge and stabilize, indicating that the model achieves balanced performance without significant overfitting or underfitting.

5. Discussion

This study demonstrates that ensemble models, particularly Random Forest, outperform other classifiers in predicting mental health conditions among remote workers. The optimized Random Forest model achieved 83% accuracy, 83% precision score, and a F1-score of 90%, with perfect recall 100%, indicating strong sensitivity. Feature engineering has significantly contributed to the model performance, highlighting hours worked, age, and years of experience as key mental health predictors.

Table 3. Comparison of Mental Health Prediction Models

Study	**Model**	**Accuracy**
(Pritam et al., 2024)	Linear Regression	65%
(Vaishnavi et al., 2022)	Stacking	81.75%
This Research	Random Forest	83%

A comparison with similiar studies highlights the effectiveness of the Random Forest model used in this research. (Pritam et al., 2024) applied a Linear Regression model and achieved an accuracy of 65%, while (Vaishnavi et al., 2022) used a K-Nearest Neighbors (KNN) classifier with a higher accuracy of 81.75%. In contrast, the Random Forest model in this study achieved 83% accuracy, slightly outperforming the existing models. This improvement demonstrates the advantage of using ensemble methods like Random Forest, which can better capture complex patterns in the data through feature randomness and decision aggregation. The result reinforces the suitability of Random Forest for mental health prediction tasks, particularly in datasets involving behavioral and psychological variables.

6. Conclusions

This study demonstrates that machine learning techniques hold considerable promise in predicting mental health risks among remote workers, with ensemble models exhibiting the most robust predictive performance, and linear models emerging as competitive and interpretable alternatives. These findings underscore the potential of data-driven approaches in supporting early identification and intervention strategies in occupational mental health, particularly within the growing population of remote workers.

However, this research is not without limitations. One of the most significant challenges encountered was the inability to access real-world mental health data due to stringent privacy regulations and ethical concerns surrounding sensitive personal information. As a result, a synthetic dataset was utilized to simulate relevant scenarios. While this approach allowed for experimentation and methodological development, it inherently limits the ecological validity of the findings. The synthetic nature of the data may introduce unknown biases, and thus, the model's performance and generalizability to actual workplace environments should be interpreted with caution. Despite these constraints, the study contributes valuable insights to the field, offering a foundational framework that can inform future research efforts. It emphasizes the importance of continued exploration into ethically sourcing anonymized real-world data, and advocates for the integration of machine learning tools into mental health monitoring systems.

Ultimately, this research highlights both the potential and the challenges of applying artificial intelligence in sensitive domains, and serves as a stepping stone toward more accurate, ethical, and impactful applications in mental health risk assessment.

Acknowledgment

We would like to express our sincere gratitude to Universitas Informatika dan Bisnis Indonesia, which has supported our research in many aspects. In particular, the Faculty of Technology and Informatics, and all the colleagues who has provided their valuable insights for this research.

References

[1] Charalampous, M., Grant, C., Tramontano, C., & Michailidis, E. "Systematically reviewing remote e-workers' well-being at work: a multidimensional approach," *European Journal of Work and Organizational Psychology*, vol. 28, pp. 51–73, 2018. doi: 10.1080/1359432X.2018.1541886

[2] Cheng, J.-P., & Haw, S.-C. "Mental Health Problems Prediction Using Machine Learning Techniques," *International Journal on Robotics, Automation and Sciences*, vol. 5, no. 2, pp. 59–72, 2023. doi: 10.33093/ijoras.2023.5.2.7

[3] Glaz, L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., & Lemey, C. "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," *Journal of Medical Internet Research*, vol. 23, 2019. doi: 10.2196/15708

[4] Idaiani, S., & Waris, L. "Depression and Psychological Stress Among Health Workers in Remote Areas in Indonesia," *Frontiers in Public Health*, vol. 10, 2022. doi: 10.3389/fpubh.2022.743053

[5] Kaushik, P., Jain, E., Gill, K. S., Upadhyay, D., & Devliyal, S. "Optimizing Mental Health Prediction by Fine-Tuning Decision Classifier Parameters for Enhanced Accuracy," *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pp. 935–939, 2024. Available: https://api.semanticscholar.org/CorpusID:271934379

[6] Lyzwinski, L.-N. "Organizational and occupational health issues with working remotely during the pandemic: a scoping review of remote work and health," *Journal of Occupational Health*, vol. 66, 2024. doi: 10.1093/joccuh/uiae005

[7] Mallick, S., & Panda, M. "Predictive Modeling of Mental Illness in Technical Workplace: A Feature Selection and Classification Approach," *2024 OPJU International Technology Conference (OTCON)*, pp. 1–7, 2024. Available: https://api.semanticscholar.org/CorpusID:272996374

[8] McKinsey. "The future of work after COVID-19," Feb. 18, 2021. Available: https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19

[9] Mohammad, B. N. S., & Siddiqui, K. "Random Forest Regressor Machine Learning Model Developed for Mental Health Prediction Based on Mhi-5, Phq-9 and Bdi Scale," *SSRN Electronic Journal*, 2021. Available: https://api.semanticscholar.org/CorpusID:237961112

[10] Nash, C., Nair, R., & Naqvi, S. "Machine Learning in ADHD and Depression Mental Health Diagnosis: A Survey," *IEEE Access*, vol. 11, pp. 86297–86317, 2023. doi: 10.1109/ACCESS.2023.3304236

[11] Nowrouzi-Kia, B., Haritos, A., Long, B.-Z. S., Atikian, C., Fiorini, L., Gohar, B., Howe, A., Li, Y., & Bani-Fatemi, A. "Remote work transition amidst COVID-19: Impacts on presenteeism, absenteeism, and worker well-being—A scoping review," *PLOS ONE*, vol. 19, 2024. doi: 10.1371/journal.pone.0307087

[12] Pritam, N., Gill, K. S., Kumar, M., Rawat, R., & Banerjee, D. "Classification of Student Mental Health Analysis using Logistic Regression and other classification techniques through Machine Learning Methods," *2024 3rd*

International Conference for Innovation in Technology (INOCON), pp. 1–5, 2024. doi: 10.1109/INOCON60754.2024.10512216

[13] Sarinastiti, N., Bimo, A., & Cole, J. "Relations of Remote Working to Mental Health," *ASPIRATION Journal*, 2022. doi: 10.56353/aspiration.v2i2.40

[14] Shatte, A., Hutchinson, D., & Teague, S. "Machine learning in mental health: a scoping review of methods and applications," *Psychological Medicine*, vol. 49, pp. 1426–1448, 2019. doi: 10.1017/S0033291719000151

[15] Sim, J., Yun, B., Park, H., Oh, J., & Yoon, J. "Rising Mental Health Issues from Remote Work Challenges in South Korea Amid COVID-19," *The European Journal of Public Health*, vol. 34, 2024. doi: 10.1093/eurpub/ckae144.2208

[16] Sofianti, T. D., Kurniawan, I., Pratama, A. T., & Florencia, J. "Measuring Worker Perception on Remote Working Adoption During COVID-19 Pandemic: An Industrial Engineering Perspective," *Engineering Science Letter*, 2023. Available: https://api.semanticscholar.org/CorpusID:258233375

[17] Vaishnavi, K., Kamath, U. N., Ashwath Rao, B., & Subba Reddy, N. V. "Predicting Mental Health Illness using Machine Learning Algorithms," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012021, 2022. doi: 10.1088/1742-6596/2161/1/012021