Jurnal Komputer, Informasi dan Teknologi Vol: 4, No 2, 2024, Page: 1-9

Analysis of the Impact of Interview-Based Feature Selection on the Performance of Machine Learning Algorithms in Mental Health Disorder Classification

Hendrick

Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara

DOI:
https://doi.org/10.53697/jkomitek.v4i2.2039

*Correspondence: Hendrick
Email:
hendrick.535210004@stu.untar.ac.id

Received: 30-10-2024
Accepted: 30-11-2024
Published: 31-12-2024

Abstract: Mental health issues in the workplace have become an increasingly important concern, particularly in the high-pressure environment of the information technology industry. This study aims to evaluate the impact of feature selection based on interviews on the performance of machine learning models in classifying mental health disorders. The dataset used is sourced from Open Sourcing Mental Illness (OSMI), which consists of various features related to employees' mental health conditions, previously used without feature selection in prior research. Through an interview with an experienced Human Capital professional with a psychological background, relevant features were selected based on domain expertise. Subsequently, machine learning models, namely Random Forest and XGBoost, were trained using two scenarios: without feature selection and with feature selection. The results of the study indicate that feature selection based on interviews can improve model accuracy by 1.67% for Random Forest and 0.67% for XGBoost. These findings emphasize the importance of integrating psychological insights into the data processing to produce more relevant and efficient models. This research provides practical contributions to assist companies in implementing early detection of mental health disorders effectively.

Keywords: Feature Selection, Mental Health, Machine learning, Random Forest, XGboost

Introduction

Mental health is an essential aspect often overlooked, particularly in the context of employees, due to high workloads and stressful work environments (Ayuningtyas et al., 2018). Mental disorders can be defined as conditions in which individuals struggle to adapt to their environment, causing unresolved issues to accumulate. This accumulation can trigger excessive stress, making individuals more vulnerable to mental health disorders. If left untreated, this condition may progress into diagnosable mental health disorders (Putri et al., 2015). Therefore, awareness and early intervention are increasingly crucial, considering the significant impact on individuals, workplace productivity, and overall well-being. Optimal mental health enables individuals to reach their maximum potential, cope with daily life pressures, work efficiently, and contribute positively to themselves and those around them (Firdausyan et al., 2023).

According to data from Indonesia's Ministry of Health, there were 826 recorded suicide cases in 2018, a relatively high number (Ardhi, 2023). Additionally, a report by the World Health Organization (WHO) indicates that 15% of the working-age population worldwide experiences mental health issues (Meilina, 2024). This data underscores the importance of early detection efforts for mental health disorders to prevent more severe consequences (Alzghoul, 2024).

Therefore, attention to employees' mental health must be enhanced, given its significant impact on workplace productivity and well-being. One innovative approach that can be implemented is the use of machine learning algorithms to classify mental health disorders using available datasets (Laksono et al., 2024). By utilizing relevant datasets, machine learning can efficiently analyze data to identify patterns and risk factors associated with mental health disorders. Leveraging existing datasets also expedites research processes as it eliminates the need for additional time and resources for data collection (Cholissodin et al., 2020). The dataset used in this study is sourced from Open Sourcing Mental Illness (OSMI), a project that collects mental health-related data from individuals working in the information technology industry (Ourbetterworld, 2019). This dataset includes various features related to mental health conditions, demographic factors, and job-related information such as job titles, mental health history, and experiences seeking professional care. The original dataset comprises 28,660 entries, covering various factors such as age, job position, and mental health history. For this study, only the first 3,000 entries are used, without filtering or random selection (Ebner, 2024).

This dataset contains 63 features representing different aspects of respondents' personal and professional lives. However, based on interviews with a Human Capital expert with a psychology background, the most relevant and significant features for classifying mental health disorders were selected, reducing the features to 46. This feature selection aims to improve the efficiency of machine learning models by focusing on variables directly contributing to workplace mental health (Garg et al., 2021). The selected features for analysis are presented in Table 1.

Table 1. Features Selected from Human Capital Interviews

Feature	Description
Tech_company	Does your company operate in the technology sector?
Tech_role	Do you have a role related to technology in your company?
MH_benefits_provided	Does your company provide mental health benefits for employees?
MH_benefits_known	Are employees in your company aware of the mental health benefits provided?
MH_discussed_formally	Is mental health formally discussed in your company?
MH_resources_offered	Does your company offer resources to support employees' mental health?
MH_anonymity_protected	Is employees' privacy protected when accessing mental health services?
MH_leave_request_response	How does the company respond to mental health-related leave requests?

MH_discussion_consequences_employer	Are there consequences for discussing mental health issues in your company?
PH_discussion_consequences_employer	Are there consequences for discussing physical health issues in your company?
MH_discussion_coworkers	Do employees feel comfortable discussing mental health issues with coworkers?
MH_discussion_supervisor	Do employees feel comfortable discussing mental health issues with supervisors?
MH_seriousness_vs_physical	Is mental health considered as serious as physical health by your company?
MH_consequences_observed	Have you observed consequences of mental health discussions at work?
Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?	Do you have health insurance (private or state-provided) covering mental health treatment?
MH_resources_known	Are employees aware of available mental health resources at work?
MH_reveal_to_coworkers	Have employees disclosed mental health conditions to coworkers?
MH_impact_on_coworkers	Does disclosure affect relationships with coworkers?
MH_productivity_affected	Has employees' productivity been affected by mental health conditions?
MH_productivity_percentage	By what percentage has productivity been affected due to mental health conditions?
PH_issue_in_interview	Have you discussed physical health issues during a job interview?
PH_issue_in_interview_reason	If yes, what was the reason for discussing physical health issues?
MH_issue_in_interview	Have you discussed mental health issues during a job interview?
MH_issue_in_interview_reason	If yes, what was the reason for discussing mental health issues?
MH_identification_impact_career	Does disclosing mental health issues affect your career?
MH_view_negatively_coworkers	Do you feel coworkers view individuals disclosing mental health issues negatively?
MH_share_with_friends_family	Do you share mental health issues with friends or family?
MH_unsupportive_response_observed	Have you observed unsupportive responses to mental health issues?
MH_influence_to_not_reveal	Is there an influence not to disclose mental health issues?
Family_history_MH	Do you have a family history of mental health issues?
Past_MH_disorder	Have you experienced mental health disorders in the past?
Current_MH_disorder	Are you currently experiencing mental health disorders?
Diagnosed_MH_conditions	Have you been diagnosed with mental health conditions?
Believed_MH_conditions	Do you believe you have specific mental health conditions?
Diagnosed_by_professional	Was your mental health condition diagnosed by a professional?
Professionally_diagnosed_conditions	Have your mental health conditions been professionally diagnosed?

Sought_treatment_from_professional	Have you sought treatment from a mental health professional?
MH_interference_when_treated_effectively	How much does mental health interfere with daily life when treated effectively?
MH_interference_when_not_treated_effectively	How much does mental health interfere with daily life when untreated effectively?
Age	What is your age?
Country_live	In which country do you live?
State_live	In which state do you live?
Country_work	In which country do you work?
State_work	In which state do you work?
Remote_work	Do you work remotely?
Work_Position	What is your profession?

This study will employ two machine learning algorithms, Random Forest and XGBoost. These algorithms are selected for their advantages in handling data with numerous features and their capability to produce accurate models through ensemble techniques (Dachi & Sitompul, 2023). Random Forest operates by constructing multiple decision trees, reducing variance to prevent overfitting, and combining their outputs through majority voting (Dachi & Sitompul, 2023; Joses et al., 2024). Meanwhile, XGBoost, a boosting algorithm, focuses on improving model accuracy by correcting errors from previous decision trees (Ompusunggu et al., 2023; Rayadin et al., 2024). The use of these two algorithms is expected to provide deeper insights into their effectiveness and precision in classifying mental health disorders within the information technology work environment. This paper aims to analyze the impact of interview-based feature selection on the performance of machine learning models in classifying mental health disorders in the workplace (Goff, 2024).

Methodology

This study aims to evaluate the impact of interview-based feature selection on the performance of machine learning models in classifying mental health disorders among employees. The research begins with the collection of a dataset sourced from Open Sourcing Mental Illness (OSMI), which includes various features related to employees' mental health conditions, such as mental health history, work environment, and job position. The next step involves feature selection through interviews with an experienced Human Capital professional with a psychology background. During these interviews, the expert provided insights into which features are deemed relevant based on practical knowledge and expertise in workplace mental health (Danielson, 2024).

After feature selection, the study compares the performance of two machine learning models, Random Forest and XGBoost, in two scenarios: without feature selection and after interview-based feature selection. The evaluation metrics used include Confusion Matrix parameters such as accuracy, precision, recall, and F1-score to measure classification effectiveness in both scenarios (Yulianti et al., 2022). The analysis is conducted to compare model performance outcomes, aiming to understand the impact of interview-based feature selection on model accuracy and efficiency.

This methodology is designed to integrate expert domain knowledge with machine learning technology, providing practical contributions to companies for early and effective detection of mental health disorders. All processes are carried out systematically to ensure reliable research outcomes (Smith, 2020). Through this approach, the study is expected to offer deeper insights into how interview-based feature selection influences the performance of machine learning algorithms in classifying mental health disorders. By selecting relevant features based on interviews with human resource professionals, the models' accuracy and efficiency in detecting mental health issues are anticipated to improve. The findings from this study are expected to contribute to developing more precise methods for workplace mental health classification and serve as a reference for researchers, practitioners, and policymakers in enhancing employee well-being through the targeted application of technology (Narciso, 2022).

Result and Discussion

To support the classification process and model evaluation, this study utilizes Python as the primary programming language due to its widespread application in data science and machine learning. Python offers various libraries that simplify data preprocessing, model training, and performance evaluation. In this research, the Pandas library is used for data manipulation and cleaning. Through Pandas, the first 3,000 records from the OSMI dataset are processed and prepared for further analysis, including converting categorical variables into numerical ones and handling missing values.

Additionally, the Scikit-learn library is employed to build and train machine learning models, specifically Random Forest and XGBoost. Scikit-learn provides a suite of algorithms and functions essential for classification and regression tasks. The train_test_split function is used to split the dataset into training and testing sets, with an 80:20 ratio. This division ensures a representative dataset and allows for an objective evaluation of the models' performance on unseen data, providing a clearer picture of their classification capabilities.

Data preprocessing also includes standardization and normalization to ensure uniform scaling of features. MinMaxScaler or StandardScaler is applied to numeric features to prevent any single feature from dominating the training process. Once the data is preprocessed, the Random Forest and XGBoost models are trained using the training dataset and evaluated with metrics such as accuracy, precision, recall, and F1-score, calculated using Scikit-learn. To enhance the clarity and informativeness of the results, visualizations are generated with Matplotlib and Seaborn, including confusion matrices that display correct and incorrect predictions for each class.

The study evaluates the models' performance under two scenarios: one using the dataset without feature selection and another using the dataset refined through interview-based feature selection with a Human Capital professional. The classification results for these scenarios are summarized in the following tables:

Table 2. Classification Results Before Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	92.33%	87.84%	96.30%	91.87%
XGBoost	98.50%	98.51%	98.15%	98.33%

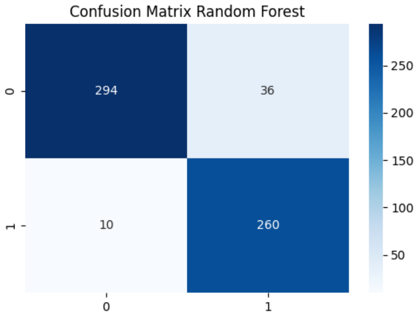

Figure 1. Confusion Matrix for Random Forest (Before Feature Selection)

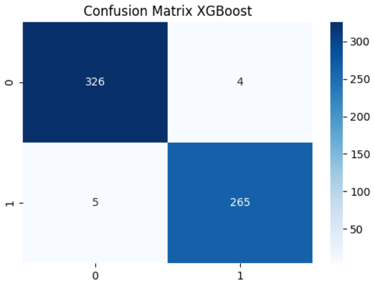

Figure 2. Confusion Matrix for XGBoost (Before Feature Selection)

Table 3. Classification Results After Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	94.00%	92.76%	95.27%	94.00%
XGBoost	99.17%	99.32%	98.99%	99.15%

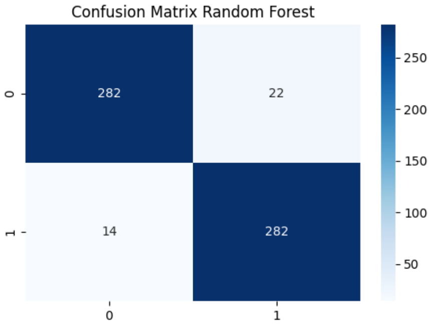

Figure 3. Confusion Matrix for Random Forest (After Feature Selection)

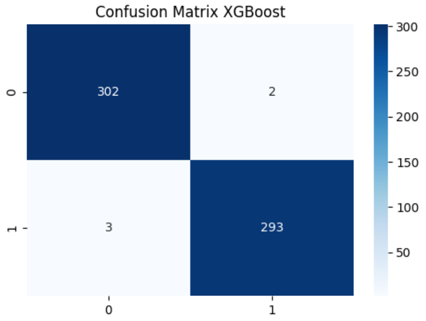

Figure 4. Confusion Matrix for XGBoost (After Feature Selection)

From the results presented in Tables 2 and 3 it is evident that models using the dataset refined through feature selection demonstrate significant performance improvements compared to those using the unfiltered dataset. These findings emphasize that interview-based feature selection, guided by a Human Capital expert with a psychology background, plays a crucial role in enhancing model accuracy and performance. Selecting relevant features based on domain knowledge allows the models to focus on factors that genuinely influence employee mental health, rather than relying on less significant attributes.

Conclusion

The results from this scenario clearly demonstrate that interview-based feature selection, conducted with a Human Capital professional with a psychology background, significantly enhances the accuracy and performance of machine learning models in classifying mental health disorders among employees. The improvements observed in accuracy, precision, recall, and F1-score highlight that models leveraging relevant and significant features can classify mental health disorders more effectively.

Specifically, the scenario utilizing interview-based feature selection shows an accuracy increase of 1.67% for the Random Forest model and 0.67% for the XGBoost model compared to the scenario without feature selection. These findings underscore that domain knowledge-driven feature selection is a highly valuable approach for developing early detection models for mental health disorders in workplace settings.

References

Alzghoul, H. (2024). Impact of Virtual Interviews on Pulmonary and Critical Care Fellowship Match An Analysis of National Data. *ATS Scholar*, *5*(1), 122–132. https://doi.org/10.34197/ats-scholar.2023-0012OC

Ardhi, S. (2023, October 13). Kementerian Kesehatan Ungkap Kasus Bunuh Diri Meningkat Hingga 826 Kasus. *Universitas Gadjah Mada*. https://ugm.ac.id/id/berita/kementerian-kesehatan-ungkap-kasus-bunuh-diri-meningkat-hingga-826-kasus/

Ayuningtyas, D., Misnaniarti, & Rayhani, M. (2018). Analisis Situasi Kesehatan Mental pada Masyarakat di Indonesia dan Strategi Penanggulangannya. *Jurnal Ilmu Kesehatan Masyarakat*, *9*(1), Article 1. https://doi.org/10.26553/jikm.2018.9.1.1-10

Cholissodin, I., Sutrisno, S., Soebroto, A. A., Hasanah, U., & Febiola, Y. I. (2020). AI, machine learning and deep learning. *Fakultas Ilmu Komputer, Universitas Brawijaya, Malang*. https://www.researchgate.net/profile/Imam-Cholissodin/publication/348003841_Buku_Ajar_AI_Machine_Learning_Deep_Learning/links/61cdd217da5d105e550a9a4a/Buku-Ajar-AI-Machine-Learning-Deep-Learning.pdf

Dachi, J. M. A. S., & Sitompul, P. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit. *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM*, *2*(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1470

Danielson, E. C. (2024). Implementation Preparation Costs of Virtual Interview Training in Pre-Employment Transition Services: A Budget Impact Analysis. *Journal of Special Education Technology*, *39*(1), 27–40. https://doi.org/10.1177/01626434231175372

Ebner, D. W. (2024). Trends in Colorectal Cancer Screening from the National Health Interview Survey: Analysis of the Impact of Different Modalities on Overall Screening Rates. *Cancer Prevention Research*, *17*(6), 275–280. https://doi.org/10.1158/1940-6207.CAPR-23-0443

M., Taqiyuddin, A., Shalahuddin, A., & Quarina, Q. (2023). Menilik Isu dan Urgensi Kesehatan Mental Pekerja Indonesia. *Jurnal Kajian*, *1*(1).

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2021). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590–1610. https://doi.org/10.1108/IJPPM-08-2020-0427

Goff, M. (2024). Investigating the impact of primary care networks on continuity of care in English general practice: Analysis of interviews with patients and clinicians from a mixed methods study. *Health Expectations*, *27*(2). https://doi.org/10.1111/hex.14032

Joses, S., Yulvida, D., & Rochimah, S. (2024). Pendekatan Metode Ensemble Learning untuk Prakiraan Cuaca menggunakan Soft Voting Classifier. *Journal of Applied Computer Science and Technology*, *5*(1), Article 1. https://doi.org/10.52158/jacost.v5i1.741

Laksono, R. D., Nurjanah, N., Sukmawati, F., Junizar, J., & Judijanto, L. (2024). *Pengantar Psikologi Umum*. PT. Green Pustaka Indonesia.

Meilina, A. D. N., Kamila. (2024, November 29). *Kemenkes Soroti Kesehatan Mental Pekerja Swasta hingga ASN, Singgung Beban Kerja—Nasional Katadata.co.id*. https://katadata.co.id/berita/nasional/6749ec4bc40bb/kemenkes-soroti-kesehatan-mental-pekerja-swasta-hingga-asn-singgung-beban-kerja

Narciso, J. (2022). How does body mass index impact self-perceived health? A pan-European analysis of the European Health Interview Survey Wave 2. *BMJ Nutrition, Prevention and Health*, *5*(2), 235–242. https://doi.org/10.1136/bmjnph-2022-000439

S., Nainggolan, A., & Sihombing, M. K. (2023). Penentuan Kelayakan Promosi Pegawai Menggunakan Algoritma Random Forest Classifier Dan Xgboost Classifier. *Jurnal Tekinkom (Teknik Informasi Dan Komputer)*, *6*(2), Article 2. https://doi.org/10.37600/tekinkom.v6i2.949

Smith, M. J. (2020). Costs of preparing to implement a virtual reality job interview training programme in a community mental health agency: A budget impact analysis. *Journal of Evaluation in Clinical Practice*, *26*(4), 1188–1195. https://doi.org/10.1111/jep.13292

Ourbetterworld. (2019). *Mental Health in Asia: The Numbers*. https://www.ourbetterworld.org/series/mental-health/support-toolkit/mental-health-asia-numbers

Putri, A. W., Wibhawa, B., & Gutama, A. S. (2015). Kesehatan mental masyarakat Indonesia (pengetahuan, dan keterbukaan masyarakat terhadap gangguan kesehatan mental). *Prosiding Penelitian Dan Pengabdian Kepada Masyarakat*, *2*(2), 252–258.

Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. *BIOS : Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), Article 2. https://doi.org/10.37148/bios.v5i2.128

Yulianti, S. E. H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, 21–26.