

Title: Using Decision Trees to Predict Turnover Intention in the Healthcare Workforce

**Name:** Nzukiso Cele

**Student Number:** 2370259

**Course Code:** FAMH4004A, Health Analytics

**Faculty:** Health Sciences

**Date of submission:** 08 September 2025

**Table of Contents**

[Introduction 2](#_Toc915273818)

[Methods 3](#_Toc1918166449)

[Data and variable description 3](#_Toc1114714246)

[Data Pre-processing 3](#_Toc139812224)

[Modelling Approach 3](#_Toc1478887159)

[Methodological Considerations 4](#_Toc741065333)

[Results 4](#_Toc133586983)

[Discussion 5](#_Toc1530263920)

[Limitations 6](#_Toc4134561)

[Implications 6](#_Toc1985179392)

[Conclusion 6](#_Toc1741699275)

[References 7](#_Toc853954564)

## **Introduction**

Workforce mental health is a critical determinant of healthcare system performance. High stress levels, frequent burnout, and low job satisfaction among healthcare workers contribute to absenteeism, decreased quality of care, and increased turnover intention (Peter *et al.,* 2024). Turnover in healthcare settings exacerbates workforce shortages, heightens workload, and undermines patient safety (Li *et al.,* 2024). Understanding predictors of turnover intention is therefore central to workforce planning and policy design.

Decision trees are machine learning algorithms that classify data by recursively partitioning predictor variables into subsets based on information gain or impurity reduction. They are particularly useful in health analytics due to their interpretability and ability to model non-linear relationships between categorical and numerical variables (Abdulqader and Abdulazeez, 2024).

This study applied a decision tree classifier to the Healthcare Workforce Mental Health Dataset obtained from Kaggle (2025). The primary objective was to predict turnover intention (Yes/No) based on a range of workplace, personal, and organisational factors. We hypothesised job satisfaction, burnout frequency, and stress levels to emerge as the strongest predictors of turnover intention, in line with findings from recent workforce research (Tolksdorf *et al.,* 2022; Li *et al.,* 2024).

## **Methods**

### Data and variable description

The dataset contained 5,000 synthetic employee records generated to reflect workplace mental health trends in healthcare (Kaggle, 2025). Each record included employee type, department, workplace factor, stress level (1–10), burnout frequency (never, occasionally, often), job satisfaction (1–5), access to employee assistance programs (EAPs), mental health absences (days), and turnover intention (yes/no). Turnover intention was the outcome variable.

Predictor selection was guided by evidence linking psychosocial factors to turnover (Tolksdorf *et al.,* 2022). Variables such as stress level, burnout frequency, and job satisfaction were expected to be primary determinants, while organisational context (department, workplace factors) and support mechanisms (EAP access) were considered secondary influences.

### Data Pre-processing

Employee ID was removed as it carried no predictive value. Categorical variables were encoded numerically to enable modeling. The dataset contained no missing values. To assess generalisability, the data was split into 80% training and 20% testing subsets, stratified to preserve class balance.

### Modelling Approach

A decision tree classifier was implemented using scikit-learn. Both Gini impurity and entropy criteria were tested, with entropy selected for the final model due to marginally superior classification performance. Overfitting was mitigated by constraining maximum depth to 5. Additionally, a simplified visualisation tree with depth limited to 3 was generated to aid interpretability.

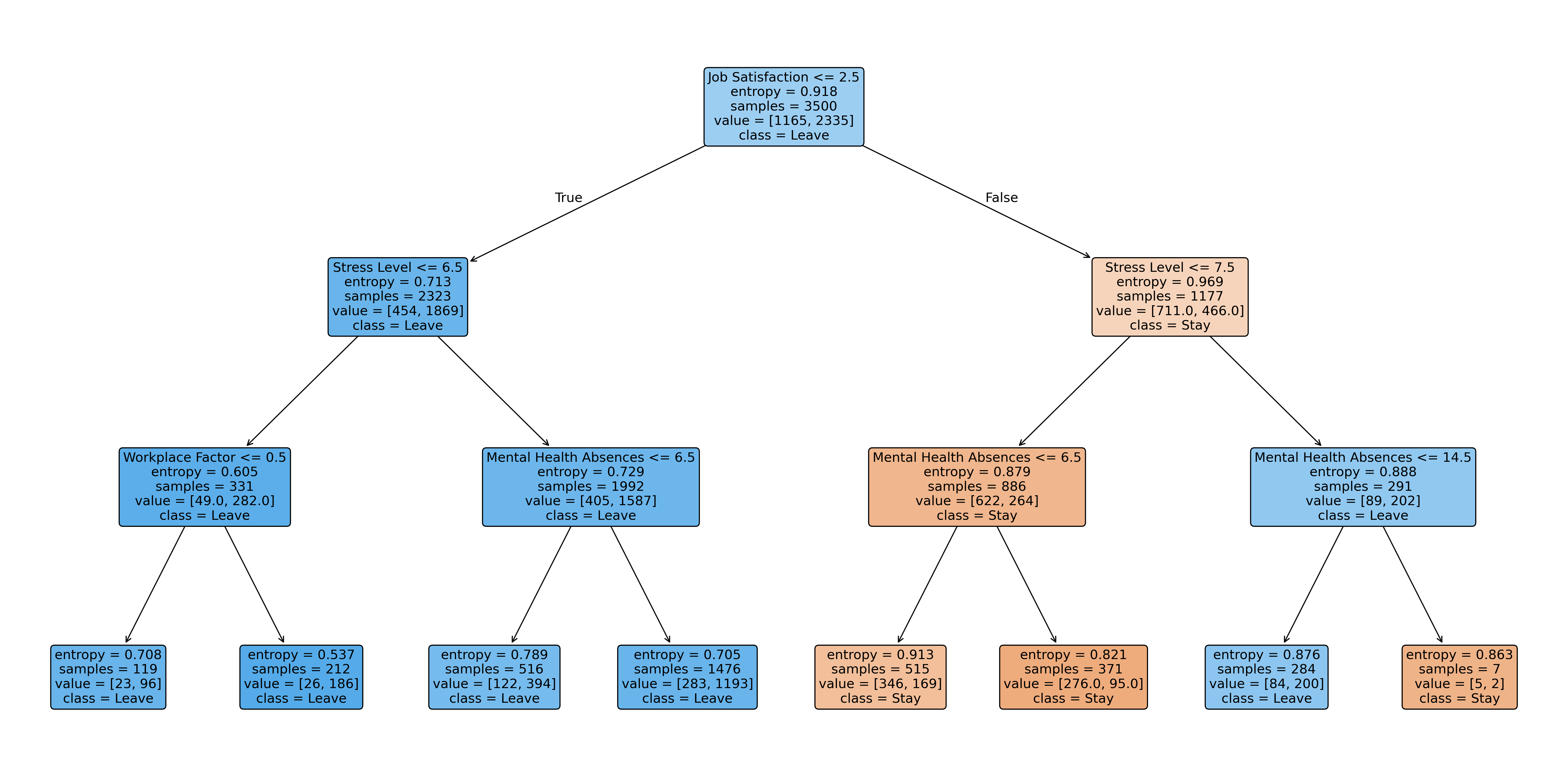
Model evaluation metrics included accuracy, precision, recall, and F1-score, supplemented by a confusion matrix. Feature importance was extracted to identify the most influential predictors.

### Methodological Considerations

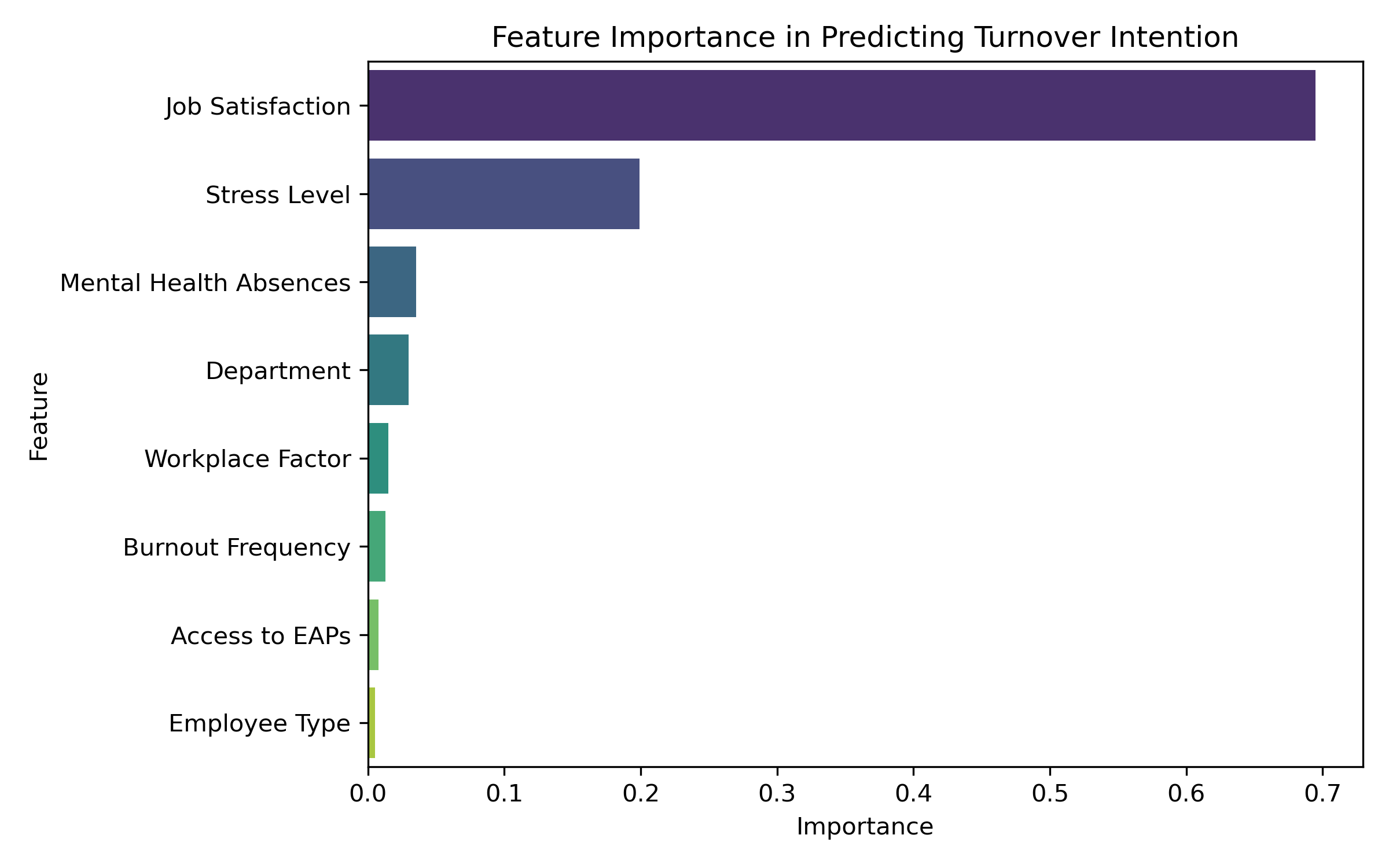
Decision trees were selected for their interpretability and capacity to model non-linear and interactional effects, which are common in psychosocial and organisational data (Srihith *et al.,* 2023). However, they are prone to overfitting and can be biased towards variables with many categories. These limitations informed the use of pruning (depth constraint) and the consideration of alternative approaches such as ensemble methods (random forests) for future work.

## **Results**

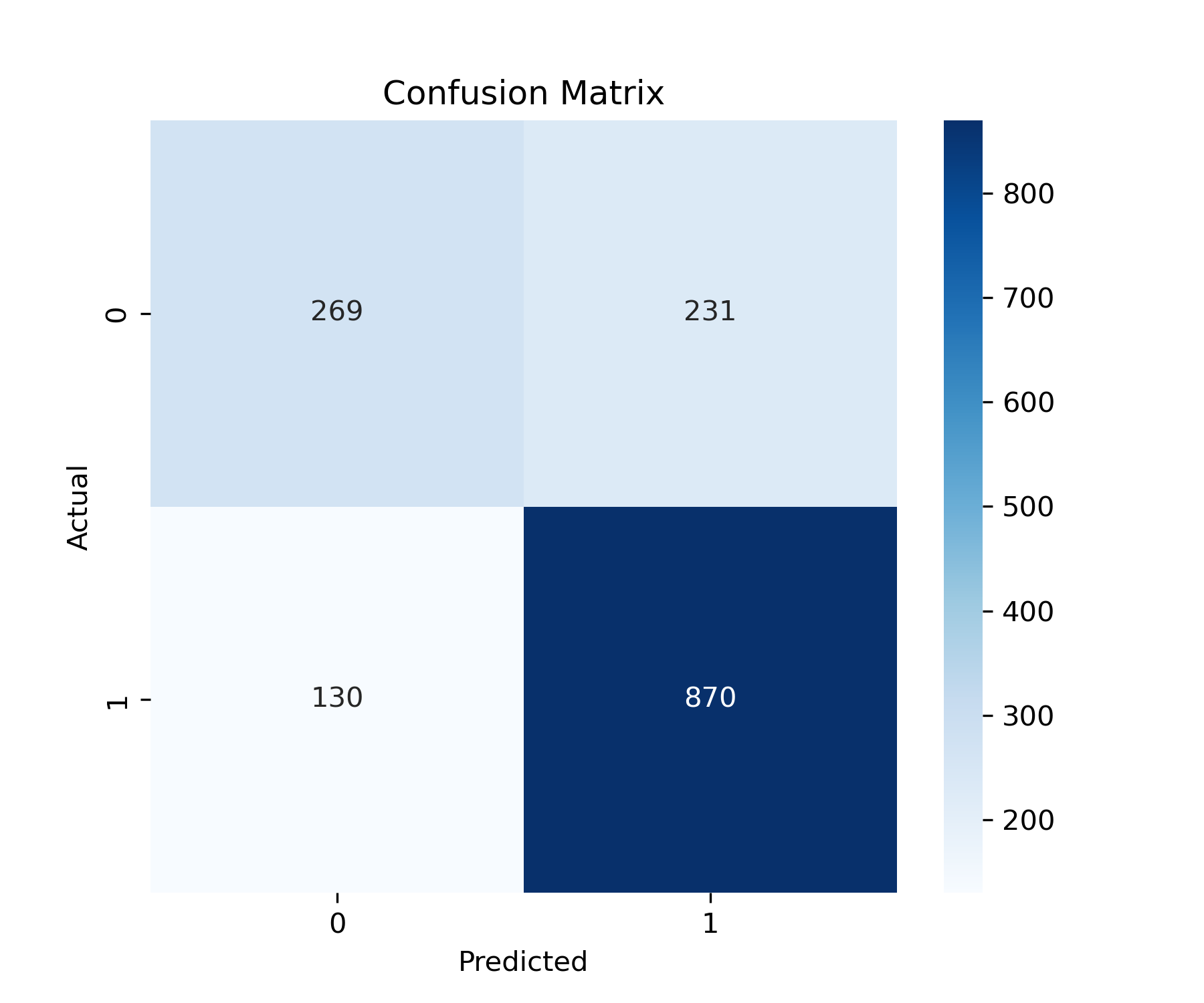
The decision tree model was evaluated using accuracy, precision, recall, and F1-score. Visualisations of the simplified tree, feature importance, and confusion matrix are presented to illustrate the model’s outputs.



**Figure 1**: Simplified Decision Tree (depth=3). This visualisation highlights the main decision rules. Employees with low job satisfaction (≤2) and high stress (≥7) were frequently classified as turnover risks.



**Figure 2**: Feature importance ranking. The ranking indicated that job satisfaction,stress level, mental health absences, and department contributed most to the model.

  
**Figure 3**: Confusion matrix. The matrix showed that most employees were correctly classified, though both false positives and false negatives occurred.

The classification report indicated accuracy: ~78%; recall for turnover intention: 81%; precision for turnover intention: 75%; and F1-score for turnover intention: 0.78.

## **Discussion**

This study showed that decision trees can effectively identify predictors of turnover intention in healthcare workforce data. The model achieved satisfactory performance, with high recall ensuring that most employees at risk of turnover were correctly identified. This is important in a workforce planning context, as failing to capture turnover risk may exacerbate staffing challenges.

For interpretability, a simplified tree (depth = 3) was used for visualisation, even though the final model was trained with depth = 5. Deeper trees often capture more complexity but result in visualisations that are cluttered and difficult to interpret. Using a simplified tree allowed clearer communication of the main classification rules while still reflecting the model’s overall structure (Abdullah *et al.,* 2022). This approach balances predictive performance with interpretability, which is particularly important in applied health analytics.

The model’s strongest predictors (job satisfaction, stress level, and burnout frequency) align with prior research. A systematic review by Tolksdorf *et al.* (2022) found that job demands, moral distress, and exhaustion significantly predicted turnover among nurses during the COVID-19 pandemic. Similarly, Li *et al.* (2024) reported that emotional exhaustion and job dissatisfaction were key determinants of turnover in primary healthcare workers. These findings reinforce the importance of addressing psychosocial stressors to reduce attrition.

Access to EAPs contributed little to predictive performance, which may reflect the dataset’s limitation of measuring only availability rather than utilisation. This aligns with Long and Cooke (2023), who argue that workplace support programs require both access and engagement to impact workforce outcomes.

## **Limitations**

First, the dataset was synthetic, reducing real-world generalisability. Second, decision trees are prone to overfitting, even with pruning. Third, categorical encoding with LabelEncoder imposed artificial order on categories, which may have influenced splits. Finally, the dataset may have imbalanced classes, requiring further methods such as resampling or class weighting.

## **Implications**

The decision tree results provide actionable insights: employees reporting low satisfaction and high stress could be prioritised for retention interventions, including workload adjustments, resilience programs, or expanded access to supportive resources (Li *et al.,* 2024). Future research should validate findings on real-world data and explore ensemble approaches (e.g., random forests) to improve predictive performance without sacrificing interpretability.

## **Conclusion**

This study applied a decision tree model to predict turnover intention in a synthetic healthcare workforce dataset. The model identified job satisfaction, stress levels, and burnout frequency as the most influential predictors, achieving an accuracy of 78% and strong recall for turnover risk. These findings align with existing evidence and highlight the value of interpretable models for workforce management.

While the synthetic nature of the data limits external validity, the study demonstrates how decision trees can uncover intuitive, rule-based insights that support decision-making in health systems. Future work should replicate the analysis with real-world data, consider ensemble methods, and assess fairness across occupational groups.

## **References**

Abdullah, T.A., Zahid, M.S.M. and Ali, W., 2021. A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. *Symmetry*, *13*(12), p.2439. Available at: [A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions](https://www.mdpi.com/2073-8994/13/12/2439) (Accessed: 06 September 2025).

Abdulqader, H.A. and Abdulazeez, A.M., 2024. A review on decision tree algorithm in healthcare applications. *The Indonesian Journal of Computer Science*, *13*(3). Available at: [A Review on Decision Tree Algorithm in Healthcare Applications | The Indonesian Journal of Computer Science](http://ijcs.net/ijcs/index.php/ijcs/article/view/4026) (Accessed: 05 September 2025).

Kaggle. 2025. *Healthcare Workforce Mental Health Dataset*. Available at: [Healthcare Workforce Mental Health Dataset](https://www.kaggle.com/datasets/rivalytics/healthcare-workforce-mental-health-dataset) (Accessed: 02 September 2025).

Li, X., Wang, J., He, L., Hu, Y., Li, C., Xie, Y., Wang, N., Luo, A. and Lu, Z., 2024. Turnover intention and influential factors among primary healthcare workers in Guangdong province, China: a cross-sectional study. *BMJ open*, *14*(11), p.e084859. Available at: [Turnover intention and influential factors among primary healthcare workers in Guangdong province, China: a cross-sectional study | BMJ Open](https://bmjopen.bmj.com/content/14/11/e084859.abstract) (Accessed: 05 September 2025).

Long, T. and Cooke, F.L., 2023. Advancing the field of employee assistance programs research and practice: A systematic review of quantitative studies and future research agenda. *Human Resource Management Review*, *33*(2), p.100941. Available at: [Advancing the field of employee assistance programs research and practice: A systematic review of quantitative studies and future research agenda - PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9664754/) (Accessed: 06 September 2025).

Peter, K.A., Voirol, C., Kunz, S., Gurtner, A., Renggli, F., Juvet, T. and Golz, C., 2024. Factors associated with health professionals’ stress reactions, job satisfaction, intention to leave and health-related outcomes in acute care, rehabilitation and psychiatric hospitals, nursing homes and home care organisations. *BMC health services research*, *24*(1), p.269. Available at: [Factors associated with health professionals’ stress reactions, job satisfaction, intention to leave and health-related outcomes in acute care, rehabilitation and psychiatric hospitals, nursing homes and home care organisations | BMC Health Services Research | Full Text](https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-024-10718-5) (Accessed: 05 September 2025)

Srihith, D., Lakshmi, P.V., Donald, A.D., Srinivas, T.A.S. and Thippanna, G., 2023. A Forest of Possibilities: Decision Trees and Beyond. *Journal of Advancement in Parallel Computing*, *6*(3), pp.29-37. Available at: [A-Forest-of-Possibilities-Decision-Trees-and-Beyond.pdf](https://www.researchgate.net/profile/T-Aditya-Srinivas/publication/374118383_A_Forest_of_Possibilities_Decision_Trees_and_Beyond/links/650ecd1282f01628f03f0dec/A-Forest-of-Possibilities-Decision-Trees-and-Beyond.pdf) (Accessed: 05 September 2025).

Tolksdorf, K.H., Tischler, U. and Heinrichs, K., 2022. Correlates of turnover intention among nursing staff in the COVID-19 pandemic: a systematic review. *BMC nursing*, *21*(1), p.174. Available at: [Correlates of turnover intention among nursing staff in the COVID-19 pandemic: a systematic review | BMC Nursing](https://link.springer.com/article/10.1186/s12912-022-00949-4) (Accessed: 05 September 2025).