**Mental Health Prediction in Tech Industry using Machine Learning**

**Abstract**

Mental health concerns in the technology sector have become increasingly prominent, drawing attention due to the industry's high-pressure environment and intense workloads. This study investigates the application of machine learning techniques to predict mental health conditions among technology industry professionals. The research utilizes a comprehensive dataset, analyzing factors such as work-life balance, job satisfaction, workplace environment, and individual well-being. The study implements and evaluates several Machine Learning approaches, including Logistic Regression, Random Forest, Decision Tree Classifier, KNN, Ada Boost Classifier, Gradient Boosting Classifier, XGB Classifier, to determine the most effective method for predicting mental health outcomes. The results highlight the potential of predictive analytics in early identification, enabling organizations to implement proactive strategies to enhance mental wellness and improve employee productivity. Additionally, this investigation emphasizes the necessity of incorporating mental health support mechanisms in technology-driven workplaces to foster a more positive work environment. The paper concludes with a discussion of the study's limitations and potential avenues for improving mental health prediction models in the future.

**1. Introduction**

**1.1 Background**

The technology sector has experienced significant growth in recent years, presenting both opportunities and challenges for its workforce. A notable concern is the increasing prevalence of mental health disorders among employees in this field. The demanding work environment, extended work hours, and constant need for innovation contribute to the psychological strain experienced by technology professionals.

Advancements in machine learning (ML) have offered promising approaches to address complex issues across various sectors, including healthcare and wellness. Specifically, the implementation of predictive models for mental health assessment has demonstrated potential in identifying at-risk individuals before their symptoms escalate. However, the application of ML methodologies to forecast mental health problems within the technology industry remains largely unexplored, with existing models often neglecting industry-specific variables and data trends.

This research aims to address this knowledge gap by examining mental health patterns in the technology sector using machine learning approaches, with a focus on identifying key indicators that could effectively predict mental health issues.

**1.2 Research Problem**

The specific challenge in predicting mental health issues in the technology industry resides in the absence of individualized analysis that considers the distinctive work conditions and stressors encountered by technology professionals. Current research and industry practices frequently fail to incorporate factors such as job role, workload, work-life balance, and job satisfaction into their predictive models, resulting in limited accuracy and generalizability.

Despite advancements in machine learning and artificial intelligence, existing methodologies do not adequately address the complexities involved in understanding mental health within the context of the technology sector. This gap leads to less effective intervention strategies and missed opportunities to support the mental well-being of technology workers.

**1.3 Objectives**

1. To evaluate the efficacy of Logistic Regression, Random Forest, Decision Tree Classifier, KNN, AdaBoost Classifier, Gradient Boosting Classifier, and XGB Classifier in predicting mental health issues among technology professionals.
2. To compare the performance of traditional machine learning methods, such as Logistic Regression, Decision Tree Classifier, and KNN, against ensemble techniques, including Random Forest, AdaBoost, Gradient Boosting, and XGB Classifier, in analyzing mental health trends within the technology industry.
3. To analyze the influence of key parameters, including job role, workload, job satisfaction, and work-life balance, on the accuracy of mental health predictions using the selected machine learning algorithms.
4. To identify the most appropriate algorithm for predicting mental health outcomes in technology workers by assessing the models based on metrics such as accuracy, precision, recall, and F1-score.

**2. Related Work**

Abdulaziz Almaleh[1] This research delves at the variables that impact workers' decisions to seek mental health care at work and makes use of preliminary data to identify obstacles. By building categorization models to forecast whether workers would seek therapy, it gives businesses valuable information about how to enhance support networks and foster a more positive work environment. Open Sourcing Mental Illness (OSMI) has produced a dataset with eight major characteristics and 27 variables. According to the study, talking about mental health concerns is more likely to have unfavorable outcomes than talking about physical health problems. It also raises the possibility of using cutting-edge machine learning techniques to improve mental health services.

Faustino Muetunda[2] The COVID-19 epidemic has made mental health problems worse and caused service interruptions. To forecast mental diseases and spot trends, machine learning classifiers such as SVM, Logistic Regression, KNN, Kernel SVM, Naive Bayes, Decision Trees, Random Forest, and ensemble approaches are employed. This study uses traditional assessment metrics to analyze these models' performance. Three datasets covering a certain age and location were gathered for the study from Psychiatric Comorbidity, Muratori et al., and Depressive Symptomatology. The results of the investigation showed that Bagging, Boosting, Bagging, KNN, and KSVM had the highest accuracy. The study also discovered that the data may be effectively used to predict and identify mental disorders using ensemble techniques, SVM, and logistic regression.

Dr.J.Arokia Renjit[3] The application of machine learning, particularly classification predictive modeling, to mental health is examined in this article. It talks about its benefits, drawbacks, and restrictions, as well as how it might be used to treat depression and other mental health conditions. Software testing and stress testing are employed in the study to enhance the models, and Flask is used to implement the optimal algorithm. Future research might link cloud models to mental health prediction and optimize the application of AI.

J.M. Imtinan Uddin[4] The study examines the likelihood of depression among Bangladeshi tech workers and focuses on risk variables. With 40% of workers in the tech sector suffering from depression and 66% reporting stress, the risk of depression is rising in this sector. The authors suggest a decision tree algorithm-based machine learning technique to estimate the likelihood of depression. They used a Google form to gather 765 responses from people aged 25 to 55. With a 98% accuracy rate, the AdaBoost decision tree algorithm fared better than the conventional method. Because depression can result in large financial losses for businesses, the study emphasizes the significance of early detection and treatment. When predicting depression risk, the AdaBoost decision tree technique performs better than previous algorithms.

Keinisha Joshi[5] Workplace productivity and quality of life are greatly impacted by mental health concerns, with stress and anxiety accounting for 52% of injuries. Deep learning and other machine learning techniques are being investigated for tracking and diagnosing mental health problems. Still, there are problems in forecasting the intensity and future events. The study assesses machine learning techniques for tracking and forecasting mental health in professionals, overcoming prior research limitations, and providing useful workplace application recommendations. The KNN method with BOW vectorization was shown to be the least effective model for comprehensive analysis of behavioral and subjective data. The top-performing models were found to be these.

Madhurima Paul[6] An analysis of mental health disorders among tech industry personnel was conducted using the Open Sourcing Mental Illness (OSMI) Mental Health in Tech Survey dataset. According to the study, 40% of respondents who had a family history of mental illness wanted medical therapy, while 12% of self-employed people sought treatment. A little over 48% of workers claim that their jobs negatively impact their mental health, while 63% of those having mental health benefits went to the doctor. Although over half of the respondents indicated their employer offered mental health benefits, over 65% said they were in need of therapy. Sixty-five percent did not know that their employer offered anonymity, and sixty-five percent sought mental health treatment. Half of the workers sought medical attention, and 62% of them felt comfortable talking to their coworkers about mental health. 80% of respondents said it was a good idea to talk about mental health with potential employers, while 40% supported discussing it with supervisors. The study identified workers with poor mental health and provided tools to support those in need using machine learning. Employers need to support staff members who are struggling with mental health issues and treat mental health with the same dignity as physical health.

Rahul Katarya[7] The dataset from the 2019 OSMI Mental Health Survey is used in this paper to detect and identify the features that contribute to mental diseases. Mental diseases are predicted using machine learning techniques like random forests and decision trees. The best classifiers are decision trees, which are followed in performance by SVM, KNN, random forest, logistic regression, and Nave Bayes. The most crucial factors for predicting a disease are family history and past mental health conditions. HR departments will ultimately profit from the application of deep learning and hybrid classifiers to enhance working conditions and mental health services for workers.

Sandhya P[8] In this study, a novel method of detecting mental health illnesses with machine learning, categorization, and genetic algorithms is presented. Using machine learning approaches, the authors searched 247 online groups and extracted psycholinguistic posts based on topics to create a combined model for recognizing traits associated with mental health. They discovered that the use of social media has raised interest in identifying red flags for mental health problems. For two-level stress identification, the suggested framework had an accuracy of 94.6%, while for multi-level stress identification, it had an accuracy of 83.4%. The study also discovered that by identifying signs of mental illnesses on social media platforms, machine learning techniques can be used to improve society. The authors identified five mental health issues, such as type 2 diabetes and depression, using eight machine learning techniques. Random forest was the best model trained for the dataset, and it included crucial features including age, gender, nationality, employment status, and family history.

Sanduni Nilushika Gamage[9] The study's main objective is to determine whether distant workers (WFH) experience mental discomfort through the use of machine learning algorithms that take into account external parameters such as sociodemographic, biological, economic, environmental, occupational, and psychological characteristics. The results showed that the CatBoost algorithm performed best in predicting distress levels, with a f1 measure of 98.5%, recall of 99.7%, precision of 97.4%, and predictive accuracy of 97.1%. In the future, the model can be expanded to include more job categories and utilized to assess different online workers in various geographic regions. IT workers with mental disorders and those who had worked from home for more than a year were not included in the study. The CatBoost algorithm outperforms other algorithms in overall measures, highlighting the importance of machine learning in automated screening of mental disorders, particularly depression, anxiety, and stress.

Siddharth Gupta[10] An analysis of mental health disorders among tech industry personnel was conducted using the Open Sourcing Mental Illness (OSMI) Mental Health in Tech Survey dataset. According to the study, 40% of respondents who had a family history of mental illness wanted medical therapy, while 12% of self-employed people sought treatment. A little over 48% of workers claim that their jobs negatively impact their mental health, while 63% of those having mental health benefits went to the doctor. Although over half of the respondents indicated their employer offered mental health benefits, over 65% said they were in need of therapy. Sixty-five percent did not know that their employer offered anonymity, and sixty-five percent sought mental health treatment. Half of the workers sought medical attention, and 62% of them felt comfortable talking to their coworkers about mental health. 80% of respondents said it was a good idea to talk about mental health with potential employers, while 40% supported discussing it with supervisors. The study identified workers with poor mental health and provided tools to support those in need using machine learning. Employers need to support staff members who are struggling with mental health issues and treat mental health with the same dignity as physical health.

**3. Methodology**

**3.1 Dataset Description**

The research makes use of several datasets about mental health in the IT sector that were obtained from Open Sourcing Mental Illness (OSMI) surveys that were carried out between 2017 and 2021. Responses from IT professionals are included in the databases, which span a range of characteristics that may have an impact on mental health outcomes.  
  
Data Sources: Data from 2017 to 2021 are included in the main dataset, which was acquired from OSMI. For analysis, the datasets were merged into a single dataframe after being read from CSV and Excel files.

Data Pre-processing:

There were multiple steps involved in data preprocessing, such as:

* Renaming Columns: To increase readability and consistency, columns have been renamed.
* Managing Missing data: We detected and eliminated from the analysis any columns that had more than 50% missing data.
* Feature Scaling: To standardize the characteristics, common scaling strategies were used.

Important Features: The dataset has attributes like:

* Age, gender, place of residence, place of employment, and status of remote work are among the demographic details.
* Factors relating to work: The extent to which a business offers benefits, how many employees work for themselves, how the employer views mental health, and the availability of mental health resources.
* Mental health indicators include past diagnosis, current state of treatment, ability to work through mental health difficulties, and degree of comfort talking about mental health with managers and coworkers.

The final dataset includes a wide range of characteristics that are essential for forecasting outcomes related to mental health in the technology sector.

**3.2 Machine Learning Models**

The study uses a number of machine learning algorithms to forecast mental health issues among IT workers. The following models were employed:  
  
1) Logistic Regression: A statistical technique called logistic regression is used to simulate the likelihood of a binary result. Because of its ease of interpretation and simplicity, it is used to forecast the possibility of mental health problems.  
  
2) Random Forest: An ensemble learning technique that boosts prediction accuracy by utilizing several decision trees. It can handle big datasets with a combination of continuous and categorical data well and is resistant to overfitting.

3) K-Nearest Neighbors (KNN): A non-parametric technique that divides data points into groups according to how close they are to one another. It is easy to implement and appropriate for smaller datasets.  
  
4) Gradient Boosting Classifier: A sophisticated ensemble technique that reduces mistakes by building models one after the other. It can manage intricate connections between variables and performs admirably with organized datasets.

5) Extreme Gradient Boosting, or XGB Classifier, is a scalable, highly effective gradient boosting method that is performance and speed optimized. It is quite popular in contests for predictive modeling and offers better accuracy for classification jobs.  
  
The reason these algorithms were chosen is that they can work with both structured and unstructured data, which is important for examining various aspects of mental health in the technology sector.

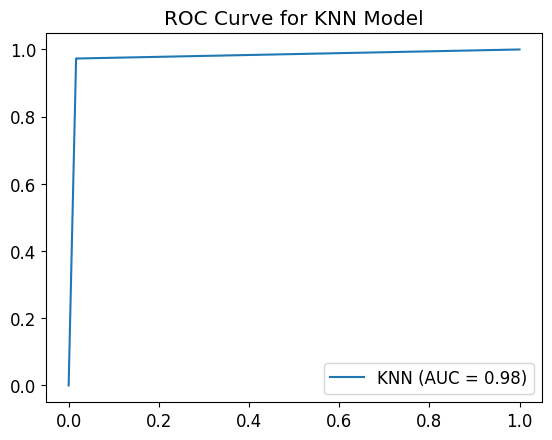
**3.3 Model Evaluation**

To guarantee precise and trustworthy predictions, the machine learning models were assessed using a variety of performance indicators. The metrics and methods listed below were applied:  
  
Cross-Validation: To make sure the models were generalizable, a k-fold cross-validation approach was used. By training the models on several subsets of the data, this technique aids in the reduction of overfitting.

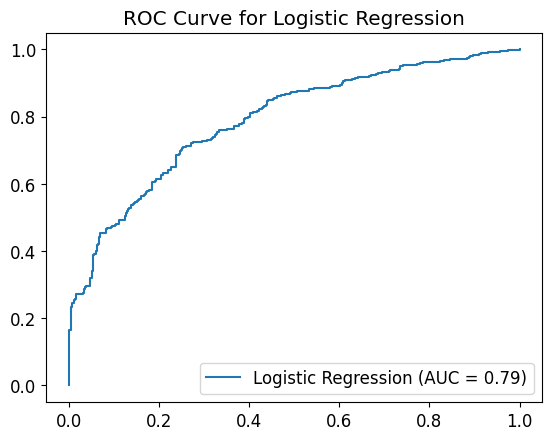
Performance Measures: The following metrics were used to evaluate the models:

* Accuracy: Calculates the percentage of accurately predicted cases among all instances.
* Precision: Measures the percentage of true positives among all positive forecasts to determine how accurate positive predictions are.
* Recall: Indicates the model's capacity to find all pertinent cases (true positives).
* F1-Score: The harmonic mean of recall and precision, which strikes a balance between the two measurements.
* The AUC-ROC curve compares the true positive rate against the false positive rate to assess how well the model can distinguish across classes.

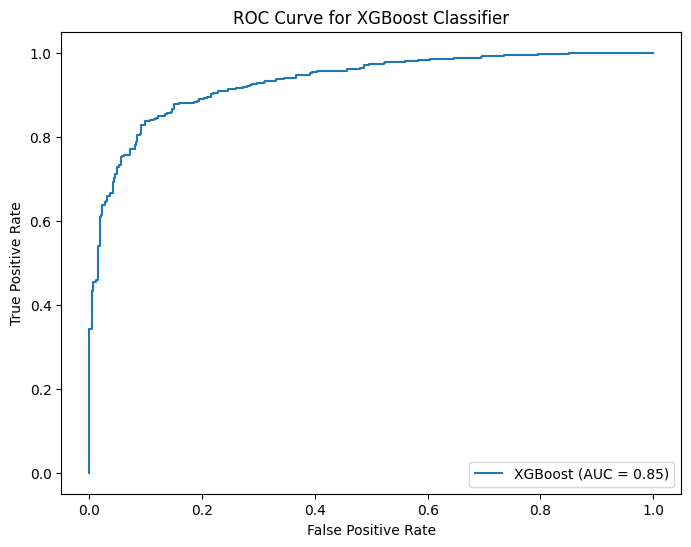
1. KNN:

  
 **Figure 1: ROC for KNN**

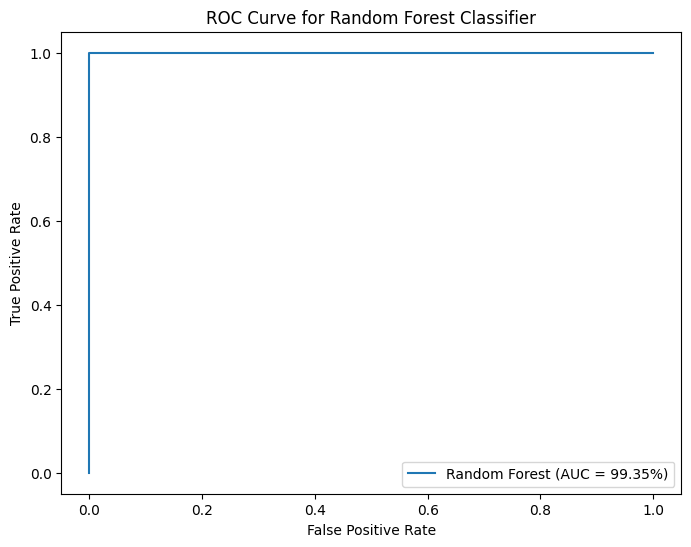
1. Logistic Regression:

  
 **Figure 2: ROC Logistic Regression**

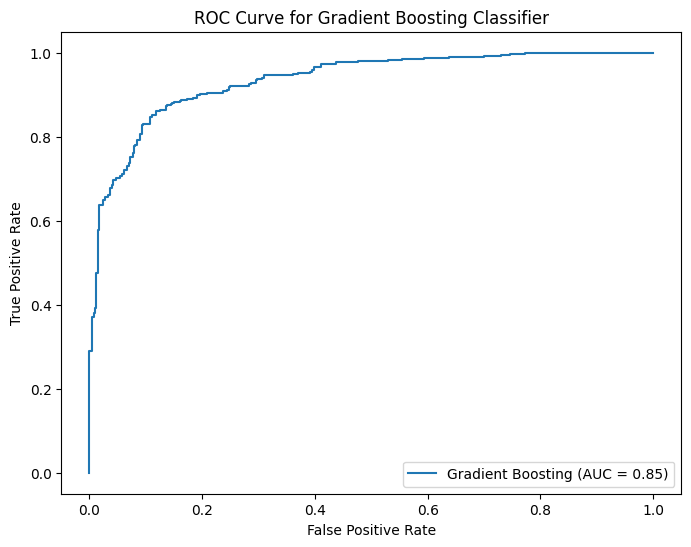
1. XG Boost:

  
 **Figure 3: ROC XG Boost**

1. Random Forest:

  
 **Figure 4: ROC for Random Forest**

1. Gradient Boosting:

  
 **Figure 5: ROC for Gradient Boosting**

**4. Results and Discussion**

**4.1 Model Performance**

The study analyzed machine learning models for predicting mental health issues in the tech industry. The Gradient Boosting Classifier and Random Forest performed well, combining multiple decision trees to capture linear and non-linear relationships. The Logistic Regression model had lower precision and recall scores, while the Decision Tree Classifier and KNN algorithms had lower accuracy due to noise and feature interactions. The results highlight the effectiveness of ensemble techniques in producing accurate and reliable predictions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| KNN | 0.95 | 0.98 | 0.97 | 97.67 |
| Logistic Regression | 0.68 | 0.52 | 0.59 | 74.49 |
| XG Boost | 0.84 | 0.69 | 0.76 | 84.52 |
| Random Forest | 1.00 | 0.98 | 0.99 | 99.35 |
| Gradient Boosting | 0.86 | 0.70 | 0.77 | 85.39 |

**4.2 Discussion**

There were notable variations in the tech industry's machine learning algorithms' ability to forecast mental health disorders. With near-perfect precision, recall, and F1-score and an astounding accuracy of 99.35%, the Random Forest model emerged as the top performer. KNN also fared well, achieving 97.67% accuracy, 0.95 precision, 0.98 recall, and 0.97 F1-score. XGBoost and Gradient Boosting, two ensemble models, produced good results with accuracy of 84.52% and 85.39%, respectively. They missed more real positive cases, nevertheless, as evidenced by their lower recall when compared to the best models. With lower precision, recall, and F1-score, the logistic regression model demonstrated the lowest accuracy, coming in at 74.49%. The findings show that complex datasets require advanced algorithms to predict mental health disorders accurately, as simpler models like logistic regression fall short in these situations.

**5.Conclusion**

The results of the research showed that machine learning models—especially ensemble methods—are quite good at predicting mental health issues in the tech sector. Upon obtaining the best accuracy, precision, recall, and F1-score, the Random Forest model was found to be the most dependable, demonstrating its capacity to manage intricate patterns within the data. Additionally, the KNN model performed well, proving that it was appropriate for this task.

While ensemble models such as XGBoost and gradient boosting performed competitively, they could not match Random Forest's robustness. These results imply that although these models are capable of handling a wide range of data, more fine-tuning or feature engineering may be necessary to increase their predictive capacity. The dataset's complexities proved too complex for the logistic regression model, despite its simplicity and interpretability. This suggests that more sophisticated methods are required to handle non-linear correlations in mental health prediction.

Overall, this study emphasizes how crucial it is to choose the right machine learning models when assessing data related to mental health, with ensemble methods like Random Forest turning out to be the most successful. In order to increase prediction accuracy and assist proactive mental health interventions in the tech industry, future research could concentrate on investigating further ensemble approaches or deep learning techniques.

**6. References**

1. Almaleh, A. (2023). *Machine Learning-Based Forecasting of mental health issues among employees in the workplace* (pp. 118–124). <https://doi.org/10.1109/iaict59002.2023.10205620>
2. Muetunda, F., Pais, S., Sabry, S., Dias, G., Pombo, N., & Cordeiro, J. (2023). *Improving mental disorder predictions using Feature-Based machine learning techninques*. <https://doi.org/10.1109/icdmw60847.2023.00165>
3. Priya, A., Garg, S., & Tigga, N. P. (2020b). Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. *Procedia Computer Science*, *167*, 1258–1267. <https://doi.org/10.1016/j.procs.2020.03.442>
4. Uddin, J. I., Fatema, K., & Dhar, P. K. (2020). Depression Risk Prediction among Tech Employees in Bangladesh using Adaboosted Decision Tree. *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. <https://doi.org/10.1109/wiecon-ece52138.2020.9397999>
5. Joshi, K., Shanna, A., Choudhary, P., Amity University Naida, Gupta, S., Minocha, S., & Amity University Naida. (n.d.). *Mental health tracking and prediction for stressed working professionals*.
6. Paul, N. M., & Das, N. S. (2023). Mental health in tech workplace: An analysis. *International Journal of Science and Research Archive*, *10*(1), 221–233. <https://doi.org/10.30574/ijsra.2023.10.1.0743>
7. Katarya, R., & Maan, S. (2020). *Predicting mental health disorders using machine learning for employees in technical and non-technical companies*. <https://doi.org/10.1109/icadee51157.2020.9368923>
8. P, S., & Kantesaria, M. (2019b). Prediction of mental disorder for employees in IT industry. In *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* (Vol. 8, Issue 6S, pp. 374–375). Blue Eyes Intelligence Engineering & Sciences Publication.
9. Gamage, S. N., & Asanka, P. P. G. D. (2022). *Machine learning approach to predict mental distress of IT workforce in remote working environments*. <https://doi.org/10.1109/scse56529.2022.9905229>
10. Kumar, V., Kochhar, A., Narula, R., Gupta, S., & Barua, P. (2023). Mental Health problem prediction of Tech Employees Using Machine Learning. *Zenodo (CERN European Organization for Nuclear Research)*. <https://doi.org/10.5281/zenodo.8337193>