ENSEMBLE LEARNING FOR PREDICTION OF MENTAL HEALTH STATUS IN EMPLOYEES: TOWARDS A STRESS-FREE WORK ENVIRONMENT

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

by

GANDLA PRAVALLIKA (20BCE1277)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2024

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I hereby declare that the thesis entitled **"Ensemble Learning for Prediction of Mental Health Status in Employees: Towards a Stress-Free Work Environment"** submitted by me, for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of Dr. Balasaraswathi V R.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: **Gandla Pravallika**

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled "**Ensemble Learning for Prediction of Mental Health Status in Employees: Towards a Stress-Free Work Environment"** is prepared and submitted by **Gandla Pravallika(20BCE1277)** to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering programme** is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr./Prof.

Date:

Signature of the Internal Examiner

Name:

Date:

Signature of the External Examiner

Name:

Date:

Approved by the Head of Department,

B.Tech. CSE

Name: Dr. Nithyanandam P

Date:

(Seal of SCOPE)

ABSTRACT

In the contemporary workplace, mental health issues among employees have become a pressing concern, impacting both individual well-being and organizational productivity. This paper presents a novel approach leveraging ensemble learning techniques for predicting the mental health status of employees. By integrating multiple machine learning algorithms, our methodology aims to provide a comprehensive assessment of stress levels among workers. Initially, we employ specific machine learning models to identify various stress indicators within the workplace environment. Subsequently, we apply ensemble techniques to refine and consolidate the outcomes from these initial models. Through this iterative process, we strive to enhance the accuracy and reliability of mental health predictions. The proposed framework offers a promising avenue for organizations to proactively address mental health challenges and foster a supportive work environment conducive to employee well-being.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. Balasaraswathi V R, Assistant Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Machine Learning.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor i/c & Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. Nithyanandam P, Head of the Department, B.Tech. CSE and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staffs at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date: **Gandla Pravallika**

TABLE OF CONTENTS

Sl.No	TITLE	PAGE No.
1.	**CHAPTER 1: INTRODUCTION**	12 – 19
	1.1 INTRODUCTION	12
	1.2 OVERVIEW OF STRESS PREDICTION	12 – 13
	1.3 OBJECTIVES	13
	1.3.1 KEY OBJECTIVES	13
	1.4 SCOPE OF THE PROJECT	13 – 14
	1.5 LITERATURE REVIEW	14 – 18
	1.6 CHALLENGES	18
	1.6.1 KEY CHALLENGES	18
	1.6.2 METHODOLOGICAL CHALLENGES	18 – 19
2.	**CHAPTER 2: THEORETICAL BACKGROUND**	20 – 26
	2.1 MACHINE LEARNING	20 – 21
	2.1.1 KNN	21 – 22
	2.1.2 RANDOM FOREST	22 – 23
	2.2 CONFUSION MATITCS	23
	2.2.1 KEY COMPONENTS OF CONFUSION MATRIX	23
	2.2.2 METRICS DERIVED FROM CONFUSION MATRIX	23
	2.3 EXSISTING SYSTEM	24

	2.3.1 COMMON APPROACHES IN EXISTING SYSTEMS	24
	2.3.2 DISADVANTAGES OF EXISTING SYSTEMS	25
	2.3.3 CHALLENGES AND LIMITATIONS OF EXISTING SYSTEMS	25 – 26
3.	**CHAPTER 3: METHODOLOGY**	27 – 33
	3.1 PROPOSED SYSTEM	27 – 29
	3.2 ADVANTAGES OF THE STRATEGY	29 – 30
	3.3 DATASET	30 – 33
4.	**CHAPTER 4: IMPLEMENTATION**	34 – 46
	4.1 DATA CLEANING	34
	4.1.1 HANDLING NULL VALUES	34 – 36
	4.2 DATA PREPOCESSING	36
	4.2.1 VISUALIZATION FOR OUTLIERS AND SKEWNESS	37 – 42
	4.3 DATA NORMALIZATION	42
	4.4 DATA RESAMPLING	43
	4.5 DATA SPLITTING	43 – 44
	4.6 KNN IMPLEMENTATION	44
	4.6.1 HYPERPARAMETER TUNING	44 – 45
	4.6.2 IMPLEMENTATION OF KNN	45
	4.7 RANDOM FOREST IMPLEMGTATION	45
	4.7.1 HYPERPARAMETER TUNING	45 – 46

	4.7.2 IMPLEMENTATION OF RANDOM FOREST	46
5.	**CHAPTER 5: RESULTS AND OBSERVATION**	47 – 50
	5.1 KNN ACCURACY	47 – 48
	5.2 RANDOM FOREST ACCURACY	48 – 49
	5.3 COMPARISION AND OBSERVATION	49 – 50
6.	**CHAPTER 6: CONCLUSION AND FUTURE WORK**	51 – 52
	6.1 CONCLUSION	51
	6.2 FUTURE WORK	51 – 52
7.	**CHAPTER 7: SOURCE CODE**	53 – 62
8.	**REFERENCES**	63 – 64

LIST OF TABLES

TABLE 1: COMPREHENSIVE OVERVIEW OF THE DATASET AND ITS ATTRIBUTES - 32

TABLE 2: HYPERPARAMETERS UTILIZED WITHIN THE RANDOM FOREST ALGORITHM. - 46

LIST OF FIGURES

FIGURE 1: PROPOSED SYSTEM ARCHITECTURE: ENSEMBLE LEARNING FOR WORKPLACE MENTAL HEALTH PREDICTION - 27

FIGURE 2: HEATMAP TO VISUALIZE MISSING VALUES IN THE DATAFRAME - 35

FIGURE 3: BOXPLOT BETWEEN AGE VS EDUCATION - 37

FIGURE 4: BOXPLOT BETWEEN HASFLEXIBLETIMINGS VS AVGDAILYHOURS - 38

FIGURE 5: DISTPLOT TO EVALUATE SKEWNESS OF JOBSATISFACTION - 39

FIGURE 6: DISTPLOT TO EVALUATE SKEWNESS OF EDUCATION - 39

FIGURE 7: DISTPLOT TO EVALUATE SKEWNESS OF AGE - 40

FIGURE 8: DISTPLOT TO EVALUATE SKEWNESS OF YEARSATCOMPANY - 40

FIGURE 9: DISTPLOT TO EVALUATE SKEWNESS OF JOBSATISFACTION - 41

FIGURE10: DISTPLOT TO EVALUATE SKEWNESS OF YEARSWITHCURRMANAGER - 41

LIST OF ACRONYMS

KNN	-	K-Nearest Neighbors
COVID-19	-	Coronavirus disease 2019
AI	-	Artificial Intelligence
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative

Chapter 1

Introduction

1.1 INTRODUCTION

In today's fast-paced and demanding work environments, the mental well-being of employees is increasingly recognized as a crucial factor influencing organizational success and individual performance. Mental health issues, particularly stress, can have profound effects on employee productivity, job satisfaction, and overall quality of life. Recognizing the significance of these challenges, this research endeavors to explore innovative approaches to predict and mitigate stress in the workplace through the application of advanced machine learning techniques.

The modern workplace presents a complex and dynamic landscape where employees are subjected to various stressors, including high workload, tight deadlines, interpersonal conflicts, and organizational changes. Consequently, addressing stress and promoting mental well-being has emerged as a priority for organizations seeking to create a supportive and conducive work environment. By leveraging data-driven methodologies such as machine learning, researchers and practitioners aim to gain deeper insights into the underlying factors contributing to stress and develop effective strategies for prevention and intervention.

Through this research, we seek to contribute to the growing body of knowledge on workplace mental health by exploring novel approaches to stress prediction and management. By harnessing the power of machine learning algorithms, we aim to develop predictive models capable of accurately assessing and forecasting stress levels among employees. Ultimately, our goal is to equip organizations with the tools and insights necessary to proactively address stress- related issues and cultivate a culture of well-being and resilience in the workplace.

1.2 OVERVIEW OF STRESS PREDICTION

Stress prediction involves the systematic analysis and forecasting of stress levels among employees within organizational settings. This process typically entails collecting and analyzing data on various individual and contextual factors that may influence stress levels, such as workload, job demands, interpersonal relationships, and organizational culture. By leveraging advanced analytical techniques, including machine learning algorithms, researchers can identify patterns, trends, and risk factors associated with stress, thereby enabling proactive interventions and support mechanisms.

The significance of stress prediction lies in its potential to inform evidence-based strategies for promoting mental well-being and reducing the negative impact of stress on individuals and organizations. By accurately identifying individuals at risk of experiencing high levels of stress, organizations can implement targeted interventions, such as stress management programs, employee assistance initiatives, and flexible work arrangements, to mitigate the adverse effects of stress and enhance overall employee resilience and performance.

However, stress prediction poses several challenges, including the subjective nature of stress perception, the complexity of stress dynamics, and the limitations of available data sources. Additionally, the effectiveness of predictive models may be influenced by various contextual factors, such as organizational culture, leadership practices, and external stressors. Addressing these challenges requires a multidisciplinary approach that integrates insights from psychology, sociology, organizational behavior, and data science to develop robust and contextually relevant predictive models.

1.3 OBJECTIVES

The primary objectives of this research are twofold: firstly, to investigate the efficacy of ensemble learning techniques, specifically KNN and Random Forest, in predicting stress levels among employees; and secondly, to identify key predictors and risk factors associated with stress in the workplace. By achieving these objectives, the aim is to enhance understanding of stress dynamics in organizational contexts and inform evidence-based interventions to promote employee well-being and foster a stress-free work environment.

1.3.1 KEY OBJECTIVES

- Evaluate effectiveness of ensemble learning techniques (KNN and Random Forest) in stress prediction.
- Identify key predictors and risk factors associated with workplace stress.
- Enhance understanding of stress dynamics in organizational settings.
- Inform evidence-based interventions for promoting employee well-being.
- Foster a stress-free work environment through targeted interventions.

1.4 SCOPE OF THE PROJECT

This research project focuses primarily on the development and evaluation of machine learning models for stress prediction in organizational settings. The scope encompasses data collection, preprocessing, model training, validation, and evaluation phases, culminating in the assessment of model performance and identification of key insights. While the research primarily targets stress prediction, the methodologies and findings

may have broader implications for addressing other aspects of mental health in the workplace. However, it is important to acknowledge the inherent limitations and constraints associated with the scope of this project, including data availability, model generalizability, and external validity considerations.

1.5 LITERATURE REVIEW

[1] Title: Machine Learning Techniques for Stress Prediction in Working Employees
Author: U Srinivasulu Reddy
Summary: This paper addresses the prevalent issue of stress disorders among working IT professionals by proposing the application of machine learning techniques for stress prediction. Using data from the OSMI mental health survey 2017, various machine learning models were trained and compared. Boosting exhibited the highest accuracy among the implemented models. Decision Trees identified significant factors influencing stress levels, including gender, family history, and availability of health benefits in the workplace. The results offer insights for industries to tailor interventions and create more comfortable work environments for employees.

[2] Title: Deep Learning Based Stress Prediction From Offline Signatures
Author: Hakan Yekta Yatbaz, Meryem Erbilek
Summary: This study presents an innovative approach to predict stress emotion from offline signature biometrics using deep learning architectures such as AlexNet, ResNet, and DenseNet. With limited research in this area, the study achieves an empirical prediction accuracy of around 77%, showcasing the potential of using signature biometrics for stress prediction.

[3] Title: Deep Learning For Time Averaged Wall Shear Stress Prediction In Left Main Coronary Bifurcations
Author: Ramtin Gharleghi, Gihan Samarasinghe, Arcot Sowmya, Susann Beier
Summary: The paper proposes a deep learning approach to estimating time average wall shear stress in coronary arteries based on vessel geometry. By leveraging deep learning, the model achieves good accuracy in predicting stress, surpassing computational simulations in efficiency. This approach enables large-scale studies for improved cardiovascular disease risk prediction.

[4] Title: A Deep Learning Approach Replacing the Finite Difference Method for In situ Stress Prediction
Author: Wenli Gao, Xinming Lu, Yanjun Peng, Liang Wu

Summary: This paper introduces a deep learning architecture, ES-Caps-FCN, for predicting in situ stress in geotechnical engineering analysis. Compared to conventional methods, the proposed deep learning approach demonstrates superior accuracy and computational efficiency, offering a promising alternative for in situ stress prediction.

[5] Title: Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health
Author: Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, Rosalind Picard
Summary: The study explores personalized machine learning models for predicting future mood, stress, and health using data from surveys, wearable sensors, smartphone logs, and weather. Multitask Learning techniques are employed to train personalized models, leading to significant performance improvements over traditional methods. The study underscores the importance of personalized approaches in mental health prediction.

[6] Title: An Integrated Multimodal Attention-Based Approach for Bank Stress Test Prediction
Author: Farid Razzak, Fei Yi, Yang Yang, Hui Xiong
Summary: This paper proposes an integrated multimodal model framework, IMBSTP, for predicting bank stress test outcomes by jointly analyzing economic conditions and banking performance profiles. The model leverages deep learning techniques and attention mechanisms to achieve improved performance compared to baseline methods, demonstrating its effectiveness in stress test prediction tasks.

[7] Title: Accurate Stress Assessment based on Functional Near Infrared Spectroscopy using Deep Learning Approach
Author: Mahya Mirbagheri, Ata Jodeiri, Naser Hakimi, Vahid Zakeri, Seyed Kamaledin Setarehdan
Summary: This study employs functional Near-Infrared Spectroscopy (fNIRS) signals and a deep learning system to assess stress induced by the Montreal Imaging Stress Task. The proposed approach achieves high accuracy in stress classification, outperforming existing methods in fNIRS studies. Its low computational cost suggests potential for real-time stress assessment applications.

[8] Title: Visual Impairment and Mental Health: Unmet Needs and Treatment Options
Author: Docia L Demmin & Steven M Silverstein

Summary: The review examines the mental health needs of individuals with vision impairment, emphasizing the significant impact of vision loss on quality of life and mental health. The paper underscores the importance of addressing mental health challenges in the visually impaired population and identifies potential treatment options to meet their unmet needs.

[9] Title: Mental Health Prevention and Promotion—A Narrative Review
Author: Vijender Singh, Akash Kumar, Snehil Gupta
Summary: This narrative review synthesizes existing literature on mental health promotion and prevention interventions, emphasizing the effectiveness of various strategies and novel approaches. Challenges and opportunities in implementing these interventions across different settings are discussed, highlighting the importance of preventive psychiatry as a public health strategy.

[10] Title: Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges
Author: Jetli Chung and Jason Teo
Summary: The paper presents a systematic review of machine learning approaches in predicting mental health problems, categorizing research articles based on mental health disorders such as schizophrenia, bipolar disorder, anxiety, depression, and posttraumatic stress disorder. Challenges and future directions for applying machine learning in mental health prediction are discussed.

[11] Title: Machine learning model to predict mental health crises from electronic health records
Author: Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros & Aleksandar Matic
Summary: The study develops a machine learning model to continuously monitor patients for the risk of mental health crises using electronic health records. The model demonstrates promising performance in predicting crises, with potential clinical utility observed in managing caseloads or mitigating crisis risks. This study contributes to proactive mental health care decision-making.

[12] Title: Prioritizing the Mental Health and Well-Being of Healthcare Workers: An Urgent Global Public Health Priority
Author: Lene E. Søvol, John A. Naslund, Antonis A. Kousoulis, Shekhar Saxena, M. Walid Qoronfleh, Christoffel Grobler, Lars Münter
Summary: The article highlights the adverse mental health impacts on healthcare workers, particularly during public health emergencies like the COVID-19 pandemic. It emphasizes the need to prioritize and protect the

mental health and well-being of healthcare workers through self-care strategies, evidence-based interventions, organizational measures, and systemic changes.

[13] Title: Methods in predictive techniques for mental health status on social media: a critical review
Author: Stevie Chancellor & Munmun De Choudhury
Summary: The systematic literature review examines predictive techniques for mental health status using social media data. The paper identifies trends and challenges in study design, data collection, preprocessing, feature selection, and model verification. Recommendations are provided to address methodological challenges and improve construct validity in future research.

[14] Title: Treatment outcomes for depression: challenges and opportunities
Author: Pim Cuijpers, Argyris Stringaris, Miranda Wolpert
Summary: The comment highlights ten key statistics related to the limitations of depression treatment outcomes, emphasizing the need for improved treatments. It discusses challenges such as the considerable proportion of patients showing improvement without treatment and the understudied ways of recovery, while also acknowledging the benefits of existing treatments.

[15] Title: Deep learning in mental health outcome research: a scoping review
Author: Chang Su, Zhenxing Xu, Jyotishman Pathak & Fei Wang
Summary: The scoping review explores applications of deep learning algorithms in mental health outcome research, categorizing relevant articles into four groups based on application scenarios. The paper discusses challenges and promising directions for using deep learning to enhance mental health diagnosis and treatment.

The literature review encompasses a diverse range of studies focusing on various aspects of mental health prediction, assessment, and treatment using machine learning and deep learning techniques. It highlights the growing interest and application of AI technologies, particularly in predicting stress, mood, and mental health outcomes. These studies demonstrate the effectiveness of machine learning models in analyzing diverse data sources, including biometrics, social media, electronic health records, and survey responses, to provide personalized insights and support for individuals' mental well-being. Additionally, the review underscores the importance of addressing mental health challenges across different populations, including working professionals, individuals with visual impairment, and healthcare workers, especially during public health emergencies like the COVID-19 pandemic. It also acknowledges the need for methodological rigor and standardized reporting in mental health prediction research to ensure the validity and reliability of predictive models and interventions. Overall, the

review emphasizes the potential of AI-driven approaches to transform mental health care delivery and promote proactive interventions tailored to individual needs, while also highlighting ongoing challenges and opportunities for future research and implementation.

1.6 CHALLENGES

Predicting stress in the workplace presents multifaceted challenges stemming from the subjective nature of stress perception, the complexity of stress dynamics, and the availability and quality of data. Firstly, the subjective nature of stress perception introduces variability among individuals in how they perceive and respond to stressors. This subjectivity complicates data collection and interpretation, as self-reported measures of stress may be influenced by social desirability bias, memory recall errors, and cultural differences. Secondly, stress is a complex phenomenon influenced by various biological, psychological, and social factors, making it challenging to capture and quantify accurately. Stress can manifest in diverse forms, including physiological symptoms, emotional distress, cognitive impairment, and behavioral changes, further complicating assessment and prediction efforts. Additionally, the availability and quality of data pose significant hurdles, as data collected from sensors, wearables, and digital platforms may vary in reliability and validity. Moreover, privacy concerns and ethical considerations surrounding the use of personal health data add layers of complexity to data collection and analysis processes.

1.6.1 KEY CHALLENGES

- Subjective nature of stress perception.
- Complexity of stress dynamics and manifestations.
- Availability and quality of data.
- Social desirability bias and cultural differences.
- Ethical considerations and privacy concerns in data collection.

1.6.2 METHODOLOGICAL CHALLENGES

In addition to the inherent complexities of predicting stress in the workplace, researchers face several methodological challenges that must be addressed to ensure the validity and reliability of predictive models. One such challenge is the selection and integration of diverse data sources, including subjective self-reports, objective biometric measurements, contextual information, and organizational data. Integrating multiple data modalities requires careful consideration of data compatibility, harmonization methods, and feature engineering techniques to extract relevant information and minimize noise and redundancy. Furthermore, establishing ground truth labels for stress levels and identifying appropriate outcome measures pose methodological challenges, particularly in the absence of universally accepted standards or gold standard assessments for stress. Variability in stress assessment tools, scales, and thresholds across studies complicates comparability and generalizability of findings, highlighting the need for standardized protocols and benchmarks in stress

prediction research. Moreover, addressing issues of data imbalance, missingness, and heterogeneity requires robust statistical methods and machine learning algorithms capable of handling complex, real-world datasets while preserving model performance and interpretability.

Overall, addressing methodological challenges in stress prediction research necessitates interdisciplinary collaboration, methodological rigor, and transparency in reporting to advance the field and facilitate reproducibility and generalizability of findings. By adopting systematic approaches to data collection, processing, modeling, and evaluation, researchers can overcome methodological hurdles and develop more accurate, reliable, and actionable predictive models for stress prediction and management in the workplace.

Chapter 2

Theoretical Background

2.1 MACHINE LEARNING

Machine learning, a subset of artificial intelligence, has emerged as a powerful tool for analyzing and extracting insights from data. Unlike traditional programming paradigms where explicit instructions are provided to solve a problem, machine learning algorithms learn from data patterns and experiences to perform tasks without being explicitly programmed. This ability to learn from data empowers machines to make predictions, decisions, and inferences, thereby automating complex tasks and enhancing efficiency across various domains.

Supervised learning is one of the fundamental paradigms in machine learning, where algorithms are trained on labeled data consisting of input-output pairs. During the training phase, the algorithm learns to map input features to corresponding output labels by optimizing a predefined objective function, such as minimizing prediction error. Once trained, the model can generalize to unseen data and make predictions for new input instances. Supervised learning algorithms include regression, where the output variable is continuous, and classification, where the output variable is categorical. For instance, in predicting employee mental health status, a supervised learning model can be trained on historical data containing features such as demographics, work environment factors, and health indicators to predict the likelihood of stress or depression.

Unsupervised learning, on the other hand, involves learning patterns and structures from unlabeled data. Without explicit output labels, unsupervised learning algorithms aim to discover hidden patterns, group similar data points, or reduce the dimensionality of the dataset. Clustering algorithms, such as k-means clustering and hierarchical clustering, are commonly used in unsupervised learning to partition data into distinct groups based on similarity or distance metrics. In the context of employee mental health prediction, unsupervised learning techniques can be applied to identify homogeneous subgroups of employees based on shared characteristics or behaviors, aiding in targeted intervention strategies or personalized support programs.

Reinforcement learning is another paradigm where agents learn to make sequential decisions through interaction with an environment to maximize cumulative rewards. Unlike supervised and unsupervised learning, reinforcement learning involves an agent taking actions in an environment and receiving feedback in the form of rewards or penalties based on the outcomes of its actions. Through trial and error, the agent learns to optimize its decision-making policy to achieve long-term objectives. Reinforcement learning has applications in various domains, including robotics, gaming, and

autonomous systems. In the context of employee mental health, reinforcement learning techniques can be employed to design adaptive interventions or feedback mechanisms that promote positive behaviors and well-being in the workplace.

In summary, machine learning techniques offer a versatile toolkit for analyzing data and deriving actionable insights in diverse domains, including employee mental health. By leveraging supervised, unsupervised, and reinforcement learning paradigms, organizations can harness the power of data-driven decision-making to enhance employee well-being, productivity, and organizational outcomes. However, it's essential to consider ethical considerations, data privacy concerns, and interpretability of machine learning models to ensure responsible and equitable deployment in real-world settings.

2.1.1 K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors (KNN) is a versatile algorithm widely used in both classification and regression tasks in machine learning. Its simplicity and effectiveness make it a popular choice for various applications, including predictive modeling in healthcare, finance, and social sciences. The core principle behind KNN is the notion of similarity: the prediction for a new data point is determined by the majority class (for classification) or the average value (for regression) of its k nearest neighbors in the feature space. In essence, KNN relies on the idea that similar data points are likely to have similar outcomes.

KNN is categorized as a non-parametric and instance-based algorithm. Non-parametric means that KNN does not make any assumptions about the underlying data distribution, making it highly flexible and adaptable to different types of datasets. Moreover, being instance-based implies that KNN stores the entire training dataset in memory, as opposed to constructing a model based on parameters. This characteristic makes KNN memory-intensive but also allows it to adapt quickly to new data points during prediction.

The selection of KNN for a project is influenced by several factors:

- Simplicity: One of the key advantages of KNN is its simplicity. The algorithm is straightforward to understand and implement, making it accessible to both beginners and experienced practitioners. Its simplicity also facilitates quick prototyping and experimentation, enabling rapid iterations in model development.
- Non-linearity: KNN is inherently capable of capturing non-linear relationships between predictors and the target variable. This flexibility is essential for modeling complex phenomena where traditional linear models may fail to capture intricate patterns in the data. In the context of stress prediction, where the relationship between predictors (e.g., demographics, work environment factors) and stress levels may be non-linear, KNN can offer more accurate predictions.

- Interpretability: KNN provides intuitive predictions based on the similarity of the new data point to its nearest neighbors. This interpretability allows stakeholders to understand and trust the model's predictions, fostering transparency and facilitating decision-making. Additionally, the simplicity of KNN's decision-making process makes it easier to communicate results to non-technical audiences.
- Robustness: KNN is robust to noisy data and outliers. By considering multiple neighboring data points during prediction, KNN can effectively handle irregularities and anomalies in the dataset. This robustness makes KNN suitable for datasets with varying levels of noise and ensures stable performance across different data conditions.

Overall, KNN offers a balance between simplicity, flexibility, and performance, making it a valuable tool for predictive modeling in a wide range of applications. However, like any algorithm, KNN has its limitations, such as computational inefficiency with large datasets and the need to choose an appropriate value for k. Understanding these trade-offs is crucial for effectively leveraging KNN in practical projects.

2.1.2 RANDOM FOREST

Random Forest stands as a powerful ensemble learning algorithm utilized for both classification and regression tasks within the realm of machine learning. This algorithm operates by constructing numerous decision trees during training and then consolidating their predictions to arrive at the final output, be it the mode of the classes (for classification) or the mean prediction (for regression). Its strength lies in the aggregation of predictions from multiple "weak learners" (individual decision trees), which collectively contribute to enhanced accuracy and generalization.

Random Forest's selection for a project is underpinned by several factors:

- Ensemble Learning: Random Forest adopts an ensemble approach by amalgamating predictions from multiple decision trees, thereby enhancing performance and mitigating the risk of overfitting. This ensemble strategy ensures that the final prediction is not overly influenced by any single decision tree, resulting in a more robust model.
- Feature Importance: An invaluable attribute of Random Forest is its capability to provide insights into feature importance. This feature analysis aids in identifying the most significant predictors influencing stress levels, thereby facilitating feature selection and enhancing model interpretability.
- Robustness to Overfitting: Random Forest exhibits inherent resilience to overfitting due to its ensemble nature. By aggregating predictions from diverse decision trees, it effectively addresses the issue of high variance, making it suitable for complex datasets with noisy features.
- Scalability: Random Forest demonstrates scalability and efficiency, particularly in handling large datasets and parallel processing. Its ability to distribute computations across multiple processors or nodes ensures

scalability, rendering it applicable to real-world scenarios with substantial data volumes.

The incorporation of both K-Nearest Neighbors (KNN) and Random Forest algorithms into the project aims to capitalize on their respective strengths. While KNN excels in capturing local patterns and relationships in the data, Random Forest offers a broader perspective by leveraging ensemble learning to aggregate predictions from diverse decision trees. This synergistic approach seeks to enhance the accuracy, robustness, and interpretability of predictive models for stress prediction in the workplace, thereby contributing to the overarching research objectives and facilitating informed decision-making within organizational contexts.

2.2 CONFUSION MATRIX

The confusion matrix is a fundamental tool in the evaluation of classification models in machine learning. It provides a comprehensive summary of the performance of a classification model by comparing predicted class labels with true class labels. The matrix consists of four quadrants: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Each cell in the confusion matrix represents the count or proportion of instances belonging to a specific combination of predicted and actual classes.

2.2.1 KEY COMPONENTS OF CONFUSION MATRIX

- True Positive (TP): Instances correctly predicted as positive by the model.
- True Negative (TN): Instances correctly predicted as negative by the model.
- False Positive (FP): Instances incorrectly predicted as positive by the model (Type I error).
- False Negative (FN): Instances incorrectly predicted as negative by the model (Type II error).

2.2.2 METRICS DERIVED FROM CONFUSION MATRIX

- Accuracy: The overall proportion of correct predictions made by the model.
- Precision: The proportion of true positive predictions among all positive predictions made by the model.
- Recall (Sensitivity): The proportion of true positive predictions among all actual positive instances in the dataset.
- Specificity: The proportion of true negative predictions among all actual negative instances in the dataset.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.

The confusion matrix serves as a valuable tool for assessing the performance of classification models, enabling researchers to identify areas of strength and weakness and fine-tune model parameters accordingly.

2.3 EXISTING SYSTEM

Existing systems for predicting stress levels in employees encompass a variety of approaches and methodologies aimed at assessing and forecasting stress within organizational settings. These systems typically employ a combination of data collection techniques, feature selection methods, and predictive models to evaluate stress dynamics. Understanding the common approaches, disadvantages, challenges, and limitations of existing systems is crucial for refining methodologies and developing more effective strategies for stress prediction and management.

2.3.1 COMMON APPROACHES IN EXISTING SYSTEMS

- Survey-Based Assessments: One prevalent approach in existing systems involves the utilization of self-reported surveys and questionnaires to gauge individuals' perceived stress levels and associated factors. These surveys often encompass various dimensions of stress, including work-related stressors, personal life stressors, coping mechanisms, and overall well-being. While survey-based assessments offer insights into individuals' subjective experiences of stress, they may be susceptible to biases and inaccuracies due to factors such as social desirability bias or recall bias.

- Biometric Sensors: Another approach involves the integration of biometric sensors, wearables, or mobile applications to collect physiological data indicative of stress. These sensors may measure parameters such as heart rate variability, skin conductance, sleep patterns, and physical activity levels. By monitoring physiological responses to stress in real-time, these systems provide objective measures of stress that complement subjective self-reports. However, challenges related to data accuracy, sensor reliability, and user compliance may impact the validity and generalizability of findings obtained through biometric sensors.

- Machine Learning Models: Many existing systems leverage machine learning algorithms to analyze collected data and predict stress levels. These models encompass a variety of techniques, including logistic regression, support vector machines, decision trees, random forests, and neural networks. Machine learning models offer the advantage of uncovering complex patterns and relationships within data, thereby enhancing the accuracy and predictive power of stress prediction systems. However, challenges such as model interpretability, overfitting, and the need for large, high-quality datasets may limit the effectiveness of machine learning approaches in certain contexts.

2.3.2 DISADVANTAGES OF EXISTING SYSTEMS

- Subjectivity: One notable disadvantage of existing systems based on self-reported surveys is their susceptibility to subjectivity and biases. Individuals' perceptions of stress may vary based on factors such as cultural background, personality traits, and situational contexts, leading to subjective assessments that may not fully capture the complexity of stress dynamics.

- Data Quality: Systems relying on biometric sensors may face challenges related to data quality and reliability. Issues such as sensor malfunction, measurement errors, and variability in physiological responses may compromise the accuracy and validity of data collected through these devices, impacting the robustness of predictive models.

- Generalizability: Another limitation of existing systems is their limited generalizability across diverse populations, organizational contexts, and cultural backgrounds. Factors such as cultural norms, work culture, and socio-economic status may influence individuals' experiences of stress, necessitating tailored approaches that account for contextual differences.

2.3.3 CHALLENGES AND LIMITATIONS OF EXISTING SYSTEMS

- Ethical Considerations: The collection and use of personal health data in existing systems raise ethical concerns related to privacy, confidentiality, and data security. Ensuring compliance with ethical guidelines and regulations, obtaining informed consent from participants, and safeguarding sensitive information are essential considerations in the development and implementation of stress prediction systems.

- Validation and Calibration: Validating and calibrating predictive models in real-world settings pose significant challenges for existing systems. Achieving high levels of accuracy, reliability, and consistency across diverse populations and environmental conditions requires rigorous validation procedures and ongoing refinement of predictive algorithms.

- Integration and Adoption: Integrating stress prediction systems into organizational settings and promoting user adoption present additional challenges. Overcoming barriers such as resistance to change, lack of

organizational support, and user skepticism requires effective communication, stakeholder engagement, and organizational buy-in.

- Multimodal Approaches: Integrating multiple data sources and modalities, such as combining self-reported surveys with biometric sensor data or leveraging contextual information from organizational records, can enrich the assessment of stress levels. By incorporating diverse perspectives and complementary sources of information, multimodal approaches offer a more comprehensive understanding of stress dynamics and enable more nuanced predictions.

- Explainable AI and Transparency: Enhancing the interpretability and transparency of machine learning models is essential for gaining stakeholders' trust and facilitating the adoption of stress prediction systems. Techniques such as explainable AI (XAI) and model-agnostic interpretability methods can provide insights into how predictive models make decisions, enabling users to understand the underlying rationale and potential biases. Transparent communication of model outputs and uncertainties fosters accountability and informed decision-making in stress management interventions.

In summary, while existing systems provide valuable insights into stress prediction and management in the workplace, they are not without limitations. Addressing challenges related to subjectivity, data quality, generalizability, ethical considerations, validation, calibration, integration, and adoption is essential for advancing the field and developing more accurate, effective, and ethically sound approaches to stress prediction and management in organizational settings. By refining existing methodologies and embracing emerging technologies and interdisciplinary collaborations, researchers can contribute to the promotion of employee well-being and organizational success in the face of stress-related challenges.

Chapter 3

Methodology

3.1 PROPOSED SYSTEM

The proposed system leverages ensemble learning techniques such as Random Forest and k-Nearest Neighbors (KNN) to predict stress levels among employees. By utilizing a combination of these algorithms, we aim to enhance the accuracy and robustness of stress prediction.

- Random Forest is a powerful ensemble method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It excels in handling large datasets with high dimensionality and provides a measure of feature importance, which can aid in understanding the factors contributing to stress levels.
- On the other hand, KNN is a non-parametric algorithm that classifies data points based on the majority class among their k-nearest neighbors. It is particularly effective for pattern recognition and classification tasks, where the decision boundaries may be complex or nonlinear.
- By combining the strengths of Random Forest and KNN, our system aims to provide accurate predictions of stress levels among employees. Additionally, once stress levels are predicted, the system can suggest measures to mitigate stress based on the identified factors contributing to stress.

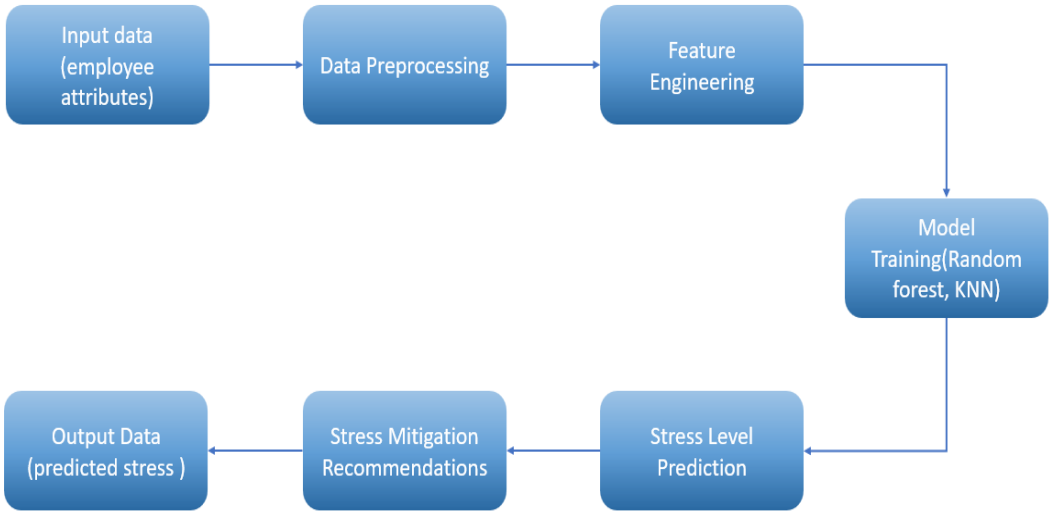

Figure 1: Proposed System Architecture: Ensemble Learning for Workplace Mental Health Prediction

The Figure 1 is the architecture of the project which includes

- Input Data:
  The input data for the proposed system consists of various attributes related to employees, such as demographic information (age, gender), job-related factors (department, job role, tenure), work environment details (workload, work hours), and potentially psychosocial factors (job satisfaction, interpersonal relationships). These attributes serve as the basis for predicting employee stress levels within the organization. Data may be collected through internal HR records, employee surveys, or other sources, ensuring relevance and reliability for stress prediction.

- Data Preprocessing:
  In the data preprocessing stage, several steps are undertaken to clean and prepare the raw input data for analysis. This involves handling missing values, outlier detection and treatment, data normalization or standardization to ensure consistency across features, and encoding categorical variables into numerical representations suitable for machine learning algorithms. Additionally, data preprocessing may involve feature scaling to bring all features to a similar scale, which is crucial for algorithms like KNN that rely on distance metrics. The goal is to create a clean, consistent dataset free from inconsistencies and errors that could adversely affect model performance.

- Feature Engineering:
  Feature engineering aims to enhance the predictive power of the model by selecting, creating, or transforming features that are most relevant to the prediction task. This may include identifying key predictors of stress based on domain knowledge or statistical analysis, creating new composite features through aggregation or interaction terms, and reducing dimensionality through techniques like principal component analysis (PCA) or feature selection methods such as recursive feature elimination (RFE). Feature engineering ensures that the model can effectively capture the underlying patterns and relationships in the data, improving its predictive accuracy and generalization performance.

- Model Training:
  In the model training phase, the preprocessed and engineered features are used to train an ensemble model comprising Random Forest and KNN algorithms. Random Forest constructs a multitude of decision trees during training and aggregates their predictions to achieve higher accuracy and robustness, while KNN relies on the similarity of neighboring data points to make predictions. By combining these two diverse algorithms, the ensemble model can effectively capture both global trends and local patterns in the

data, leading to improved stress level predictions. Model training involves splitting the dataset into training and validation sets, tuning hyperparameters, and evaluating model performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

- Stress Level Prediction:
  Once trained, the ensemble model is capable of predicting stress levels for new employees based on their attribute profiles. Given a set of employee attributes as input, the model calculates the predicted stress level, providing valuable insights into potential stressors and risk factors within the organization. This prediction enables proactive intervention and support measures to be implemented to mitigate stress and promote employee well-being, contributing to a healthier and more productive work environment.

- Stress Mitigation Recommendations:
  Based on the predicted stress levels and the factors contributing to stress identified by the ensemble model, personalized recommendations for stress mitigation strategies can be generated. These recommendations may include tailored interventions such as stress management workshops, mindfulness training, workload adjustments, flexible work arrangements, or counseling services. By providing targeted recommendations aligned with individual needs and organizational context, the system facilitates proactive stress management and fosters a culture of employee support and well-being.

- Output Data:
  The output data generated by the system consists of predicted stress levels for employees, along with any associated recommendations for stress management. This information can be used by HR professionals, managers, and organizational leaders to implement targeted interventions, monitor trends in employee stress levels over time, and evaluate the effectiveness of stress mitigation strategies. Ultimately, the output data serves as a valuable resource for promoting employee health, resilience, and productivity within the organization.

3.2 ADVANTAGES OF THE STRATEGY

The proposed strategy offers several advantages:

- Improved Prediction Accuracy: Ensemble learning techniques like Random Forest and KNN often yield higher prediction accuracy compared to individual models. By aggregating predictions from multiple models, the system can capture diverse patterns and reduce the risk of overfitting.

- Interpretability: Random Forest provides a measure of feature importance, allowing us to interpret the factors influencing stress levels among employees. This insight can inform decision-making processes related to stress management interventions and organizational policies.
- Flexibility: Ensemble learning techniques are flexible and can accommodate various types of data and problem domains. Whether the dataset exhibits linear or nonlinear relationships, these techniques can adapt and provide reliable predictions.
- Robustness: Ensemble methods are robust to noisy data and outliers. By combining predictions from multiple models, the system can mitigate the impact of individual model biases or inaccuracies.
- Scalability: Both Random Forest and KNN are scalable algorithms, capable of handling large datasets efficiently. This scalability ensures that the proposed system can be applied to organizations of different sizes without significant computational overhead.

Overall, the proposed strategy offers a comprehensive approach to predicting stress levels among employees and recommending appropriate interventions, thereby promoting employee well-being and organizational effectiveness.

3.3 DATASET

The dataset encompasses a diverse array of 30 attributes collected from employees within a workplace setting, aiming to comprehensively capture various facets of their demographics, work-related dynamics, and personal characteristics. This holistic approach includes not only traditional work-related attributes but also personality-related factors outside the workplace, contributing to a more nuanced and accurate assessment of employees' mental health status.

Sl.No	**Attribute**	**Description**
1.	EmployeeID	Each employee is uniquely identified within the dataset, facilitating individual-level analysis and tracking.
2.	Target	The target variable denotes the mental health status of employees, crucial for predictive modeling and intervention strategies.
3.	Age	Providing insights into the age distribution within the workforce, this attribute offers context regarding potential age-related factors influencing mental well-being.
4.	AvgDailyHours	Reflecting the average daily working hours, this attribute sheds light on the workload and potential stressors faced by employees.
5.	Department	Categorizing employees based on their department enables exploration of potential differences in stress levels across organizational units.

6.	Education	The highest level of education attained by employees serves as a proxy for educational background and skillset diversity.
7.	EducationField	Offering insight into employees' academic specialties, this attribute captures variations in knowledge domains and professional expertise.
8.	Gender	Considering gender diversity is essential for understanding potential disparities in stress experiences and coping mechanisms.
9.	HasFlexibleTimings	This binary attribute indicates whether employees have flexibility in their work schedules, impacting work-life balance and stress management.
10.	IsIndividualContributor	Distinguishing between individual contributors and team members aids in examining stressors unique to different roles and responsibilities.
11.	JobInvolvement	Assessing the level of engagement in job-related tasks provides insights into employees' commitment and potential sources of job-related stress.
12.	JobRole	Categorizing employees based on their roles facilitates role-specific analysis and identification of stressors inherent to each position.
13.	JobSatisfaction	Reflecting employees' satisfaction with their roles and responsibilities, this attribute influences overall job-related well-being.
14.	LeavesTaken	The number of leaves taken by employees indicates their utilization of time-off and potential stress-related absences.
15.	MaritalStatus	Exploring marital status offers insights into the intersection of personal life and work-related stressors.
16.	MicromanagedAtWork	Assessing perceptions of micromanagement provides context regarding autonomy and job control, influencing stress levels.
17.	MonthlyIncome	Employees' monthly income serves as a proxy for socioeconomic status, which may impact access to resources for stress management.
18.	NumCompaniesWorked	The number of companies employees have worked for in the past may indicate career stability and adaptability to organizational changes.
19.	PercentSalaryHike	Reflecting recent changes in salary, this attribute may influence job satisfaction and overall well-being.
20.	PerformanceRating	Employees' performance ratings offer insights into their professional competence and potential sources of job-related stress.

21.	RelationshipSatisfaction	Assessing satisfaction with personal and professional relationships outside the workplace is crucial for understanding holistic well-being.
22.	RemoteWorkSatisfaction	Satisfaction with remote work arrangements, if applicable, influences work-life balance and stress management strategies.
23.	SelfMotivationLevel	Evaluating employees' self-motivation levels provides insights into their resilience and coping mechanisms in the face of stress.
24.	TotalWorkingYears	Reflecting cumulative work experience, this attribute offers context regarding career trajectory and potential stressors associated with tenure.
25.	TrainingTimesLastYear	The number of training sessions attended by employees reflects organizational investment in skill development and potential stressors associated with learning.
26.	WorkLifeBalance	Perceptions of work-life balance influence overall well-being and job satisfaction, impacting stress levels.
27.	WorkLoadLevel	Employees perceptions of workload provide insights into task demands and potential sources of stress.
28.	YearsAtCompany	The duration of employees' tenure at the current company may influence organizational belongingness and stress levels associated with job stability.
29.	YearsSinceLastPromotion	The time elapsed since employees' last promotion may impact job satisfaction and career progression-related stress.
30.	YearsWithCurrManager	The duration of employees' working relationship with their current manager influences supervisory support and potential stressors associated with leadership dynamics.

Table 1: Comprehensive Overview of the dataset and its attributes

By incorporating a diverse range of attributes, as detailed in Table 1, including both work-related and personality-related factors outside the workplace, this dataset facilitates a holistic understanding of employee mental health, enabling more accurate prediction and targeted interventions for fostering a supportive and stress-free work environment.

The comprehensive dataset outlined in Table 1 represents a significant advancement in understanding and addressing employee mental health within organizational settings. By encompassing a diverse array of 30 attributes, ranging from demographic factors to work-related dynamics and personal characteristics, the dataset offers a nuanced perspective on the multifaceted nature of stress and well-being in the workplace. This holistic approach acknowledges that mental health outcomes are influenced by a complex interplay of individual traits, organizational factors, and socio-environmental contexts, underscoring the importance of considering a wide range of variables in stress prediction and management efforts.

Moreover, the inclusion of personality-related factors outside the workplace in the dataset enriches the predictive modeling process by capturing aspects of employees' lives that may impact their mental health beyond the confines of their professional roles. By recognizing the interconnectedness of personal and work-related experiences, organizations can develop more tailored and effective interventions to support employees' well-being. Through advanced analytical techniques such as machine learning, researchers can leverage this rich dataset to uncover hidden patterns, identify key predictors of stress, and develop targeted strategies for mitigating stressors and promoting resilience among employees. Ultimately, the utilization of such comprehensive datasets holds great promise for fostering a culture of well-being and empowerment in the workplace, where employees feel supported, valued, and equipped to thrive amidst the challenges of modern work environments.

Chapter 4

Implementation

4.1 DATA CLEANING

Data cleaning is an essential pre-processing step in any data analysis or predictive modeling project. Null values, also known as missing data, can arise due to various reasons such as data entry errors, incomplete records, or intentional omission. Handling null values is crucial as they can skew statistical analyses and adversely impact the performance of machine learning algorithms. In the context of predicting employee mental health status, addressing null values is paramount to ensure the accuracy and reliability of the predictive model.

One approach to handling null values is imputation, where missing values are replaced with estimated values based on the available data. Common imputation techniques include mean imputation, median imputation, or mode imputation, where the mean, median, or mode of the respective attribute is used to fill in missing values. Another approach is to remove observations or attributes with a high proportion of missing values if they do not contribute significantly to the predictive task. However, caution must be exercised to avoid losing valuable information by indiscriminate removal of data.

Additionally, advanced imputation methods such as predictive modeling-based imputation or k-nearest neighbors imputation can be employed to impute missing values based on relationships with other variables in the dataset. These methods utilize machine learning algorithms to predict missing values based on patterns observed in the existing data. However, it's essential to evaluate the performance of imputation techniques and consider the potential impact on the downstream analysis. Robust data cleaning procedures ensure the integrity and quality of the dataset, laying a solid foundation for accurate predictive modeling and meaningful insights into employee mental health status.

4.1.1 HANDLING NULL VALUES

The data cleaning process is critical for ensuring the integrity and quality of the dataset used for predictive modeling. Null values, or missing data, are common in real-world datasets and can significantly affect the performance of machine learning algorithms if left unaddressed. In this project aimed at predicting employee mental health status, the first step involved identifying and handling null values within the dataset.

To begin with, we conducted a comprehensive assessment of the dataset to identify the presence of null values across various attributes. This involved using descriptive statistics to quantify the extent of missingness within each attribute. Visualizations such as heatmaps were also utilized to provide a graphical representation of missing values, facilitating a deeper understanding of their distribution across the dataset.

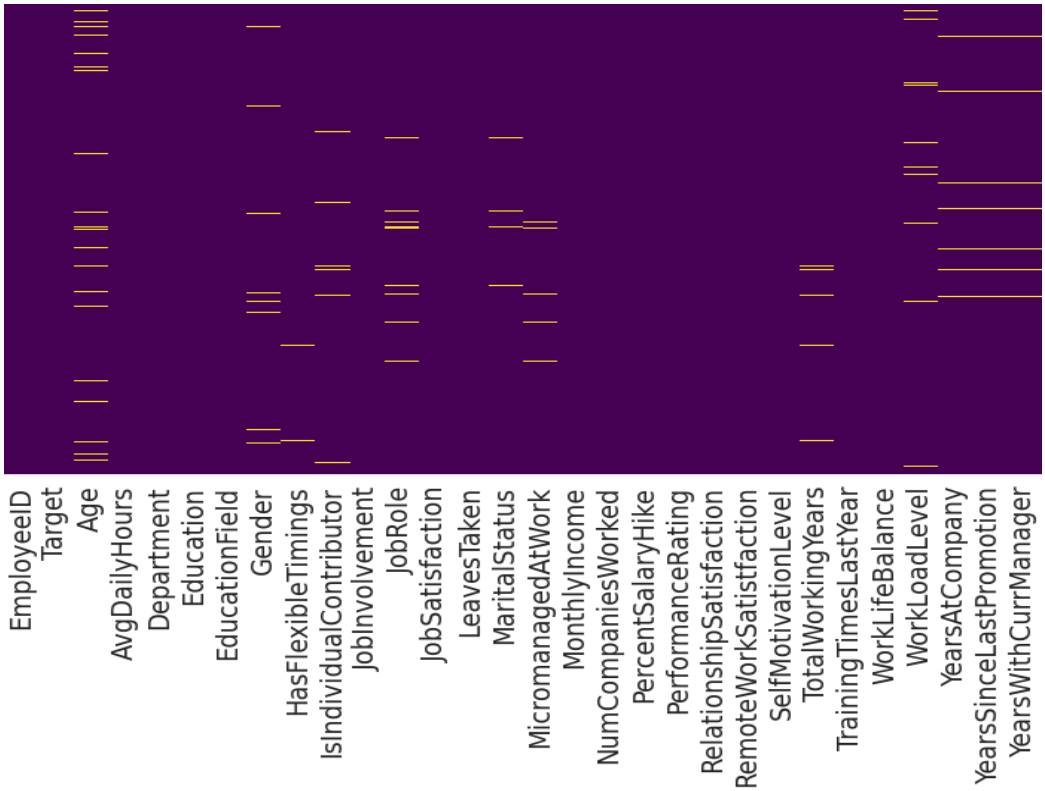

Figure 2: Heatmap to visualize missing values in the DataFrame

Once null values were identified as shown in Figure 2, we employed appropriate strategies to handle them effectively. For attributes with a small proportion of missing values, we opted for imputation techniques such as mean or median imputation, where missing values were replaced with the mean or median of the corresponding attribute. This approach helps preserve the overall distribution of the data while filling in missing values with reasonable estimates.

In cases where attributes had a significant proportion of missing values or where imputation was not feasible, we made the decision to remove those attributes entirely from the dataset. This ensured that the remaining data remained consistent and reliable for subsequent analysis and modeling tasks. Additionally, we carefully considered the implications of missing values on the interpretability and generalizability of the predictive models, taking steps to minimize their impact wherever possible.

Overall, the data cleaning process was conducted systematically and rigorously to address null values and ensure the quality and completeness of the dataset. By implementing appropriate handling strategies, we were able to mitigate potential biases and inaccuracies in the analysis, laying a solid foundation for subsequent data preprocessing and modeling tasks.

4.2 DATA PREPROCESSING

Data preprocessing plays a crucial role in preparing the dataset for analysis and modeling by addressing issues such as outliers and skewness. Outliers can significantly affect the performance of predictive models by distorting statistical measures and influencing model fitting. Therefore, identifying and appropriately handling outliers is essential to ensure the robustness and reliability of the predictive model. In addition to visualizing outliers using box plots, other techniques such as Z-score normalization or Winsorization can be employed to mitigate their impact. Z-score normalization involves standardizing the distribution of data by scaling each feature to have a mean of 0 and a standard deviation of 1, thereby reducing the influence of outliers on the analysis. Winsorization, on the other hand, involves replacing extreme values with less extreme values within a predetermined range, effectively reducing the impact of outliers while preserving the overall distribution of the data.

Skewness in the distribution of data can also pose challenges in predictive modeling, as it violates the assumptions of normality underlying many statistical techniques. Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Addressing skewness can involve transformations such as logarithmic transformation or power transformation to make the distribution more symmetrical. Logarithmic transformation is commonly used for positively skewed data, while power transformations such as square root or cube root transformation can be applied to data exhibiting different degrees of skewness. By visualizing the distribution of key features and assessing skewness, appropriate transformations can be applied to achieve a more balanced and representative dataset for modeling.

Furthermore, it's essential to consider the context of the data and the specific requirements of the predictive task when deciding on preprocessing techniques. While outliers and skewness may need to be addressed to improve model performance, it's crucial to strike a balance between data manipulation and preserving the integrity of the original data. Thorough data preprocessing ensures that the dataset is well-suited for modeling and facilitates the generation of accurate insights into employee mental health status.

4.2.1 VISUALIZATION FOR OUTLIERS AND SKEWNESS

Before proceeding with data preprocessing, it was essential to gain insights into the distribution of key features and identify potential anomalies such as outliers and skewness. Outliers are data points that deviate significantly from the rest of the dataset and can adversely affect the performance of predictive models if left unaddressed. Skewness, on the other hand, refers to asymmetry in the distribution of data points and can impact the assumptions underlying many statistical models.

To visualize outliers, we employed box plots as shown in Figure 3 and Figure 4, which provide a visual representation of the distribution of data and highlight any extreme values that may be present. By examining box plots for key features related to employee attributes and mental health indicators, we were able to identify potential outliers that warranted further investigation. Outliers can arise due to various reasons, including data entry errors, measurement inaccuracies, or genuine anomalies in the underlying data distribution.

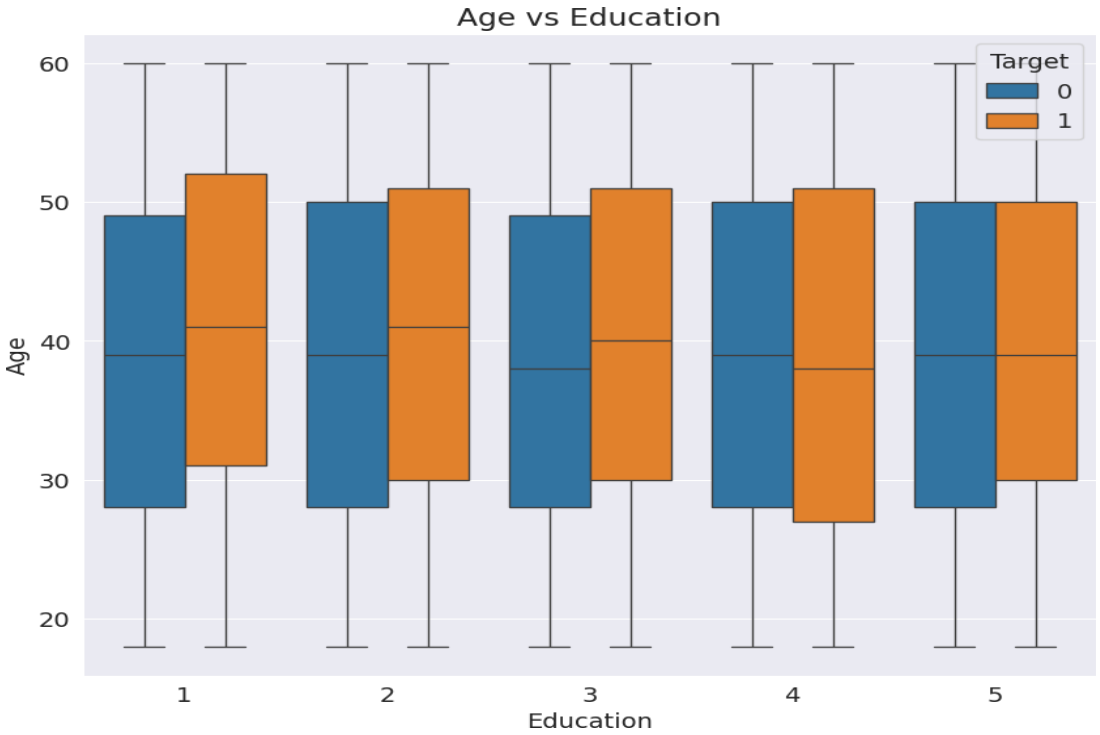

Figure 3: Boxplot between Age vs Education

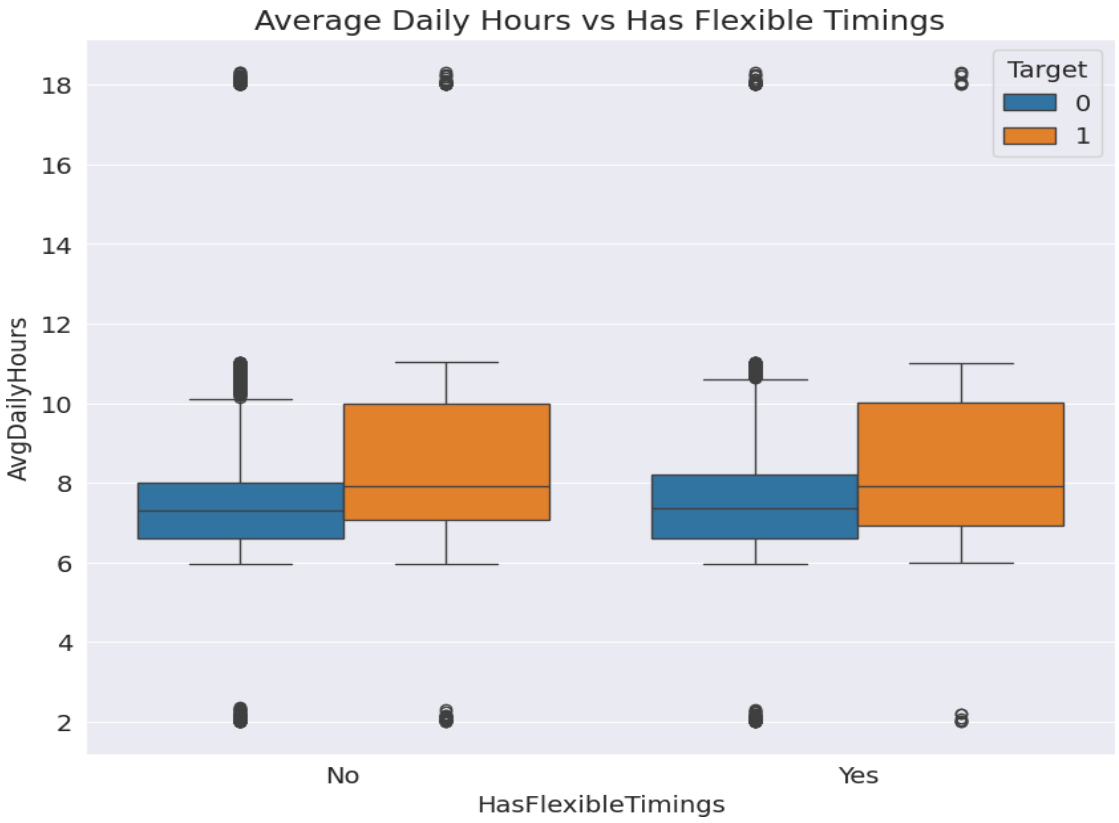

Figure 4: Boxplot between HasFlexibleTimings vs AvgDailyHours

In addition to outliers, we also assessed the skewness of key features using distribution plots (distplots) as shown in Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, which visualize the distribution of data points along with their probability density. Skewed distributions can indicate departures from normality and may require appropriate transformations to achieve symmetry or to normalize the data distribution. By visualizing distplots for relevant features, we gained insights into the shape of the data distribution and identified potential skewness that needed to be addressed during preprocessing.

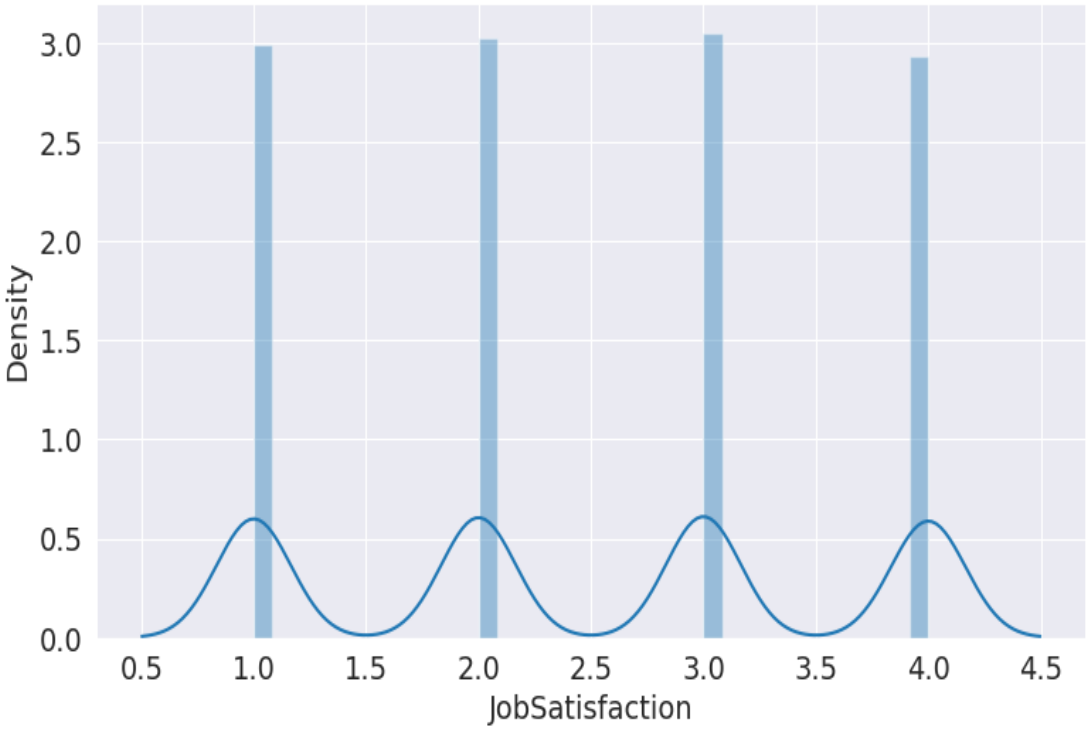

Figure 5: Distplot to Evaluate Skewness of JobSatisfaction

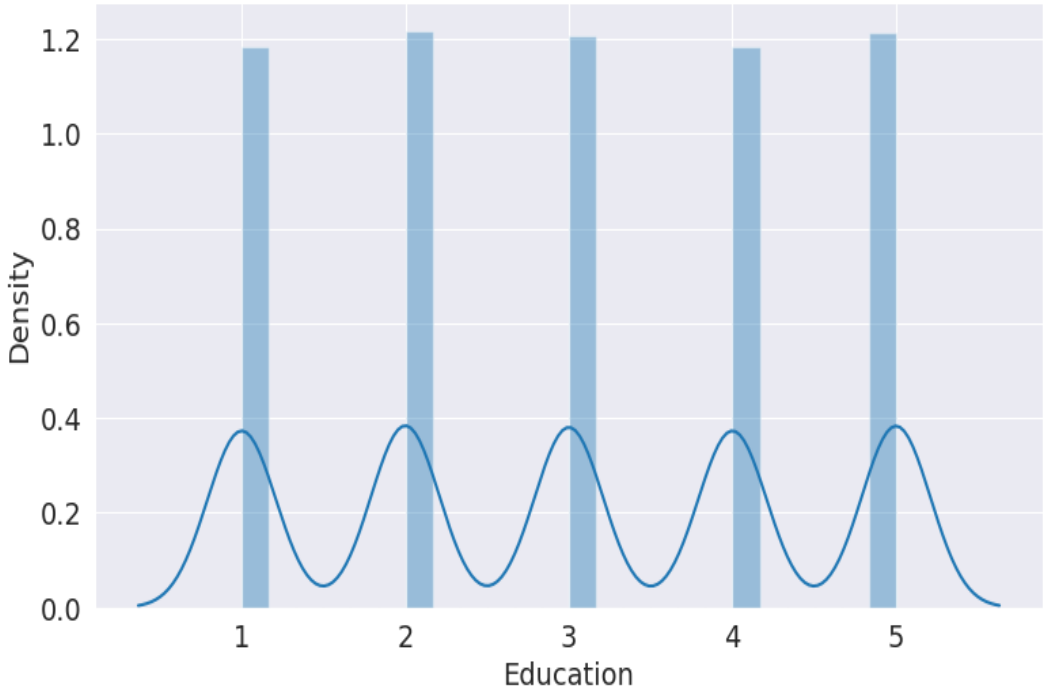

Figure 6: Distplot to Evaluate Skewness of Education

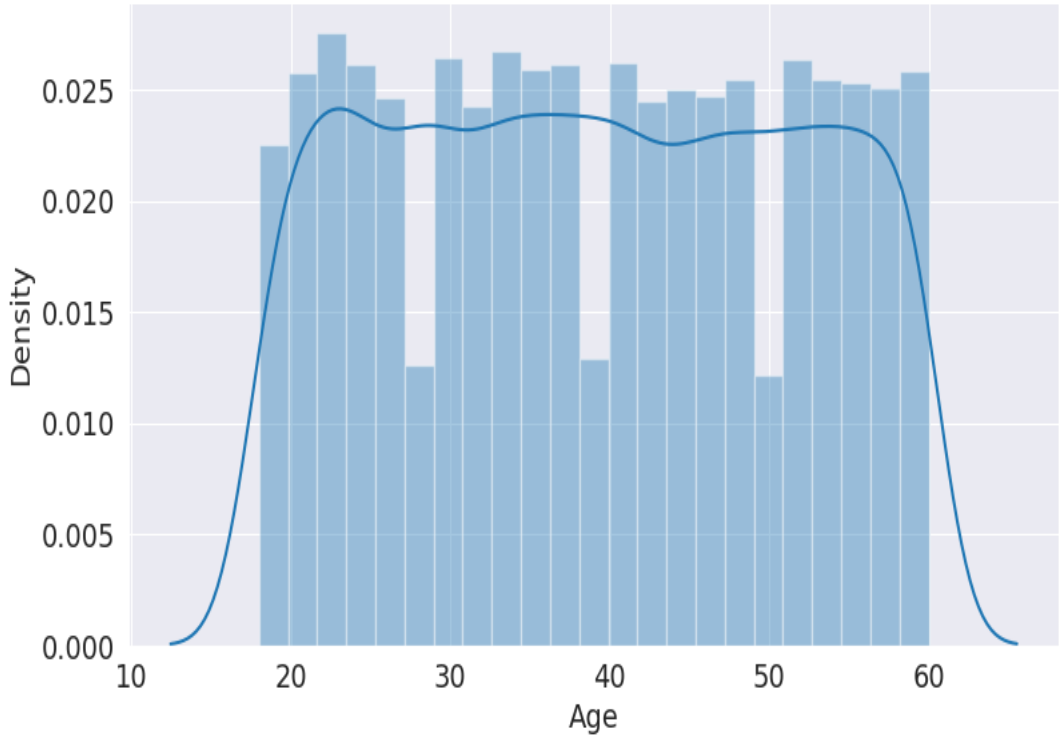

Figure 7: Distplot to Evaluate Skewness of Age

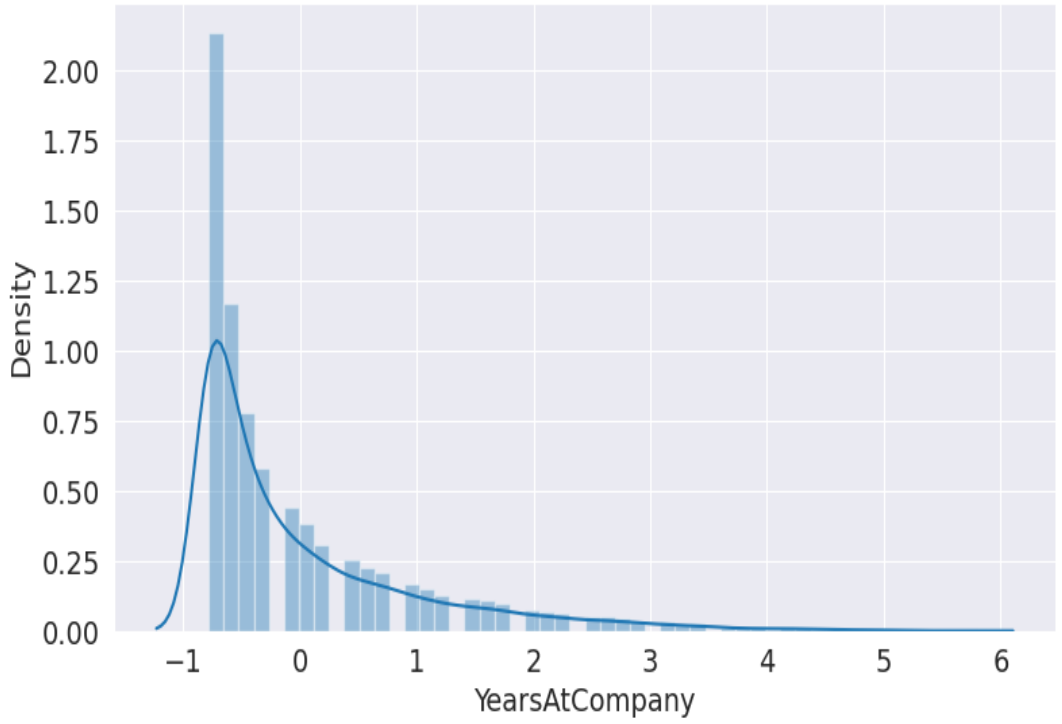

Figure 8: Distplot to Evaluate Skewness of YearsAtCompany

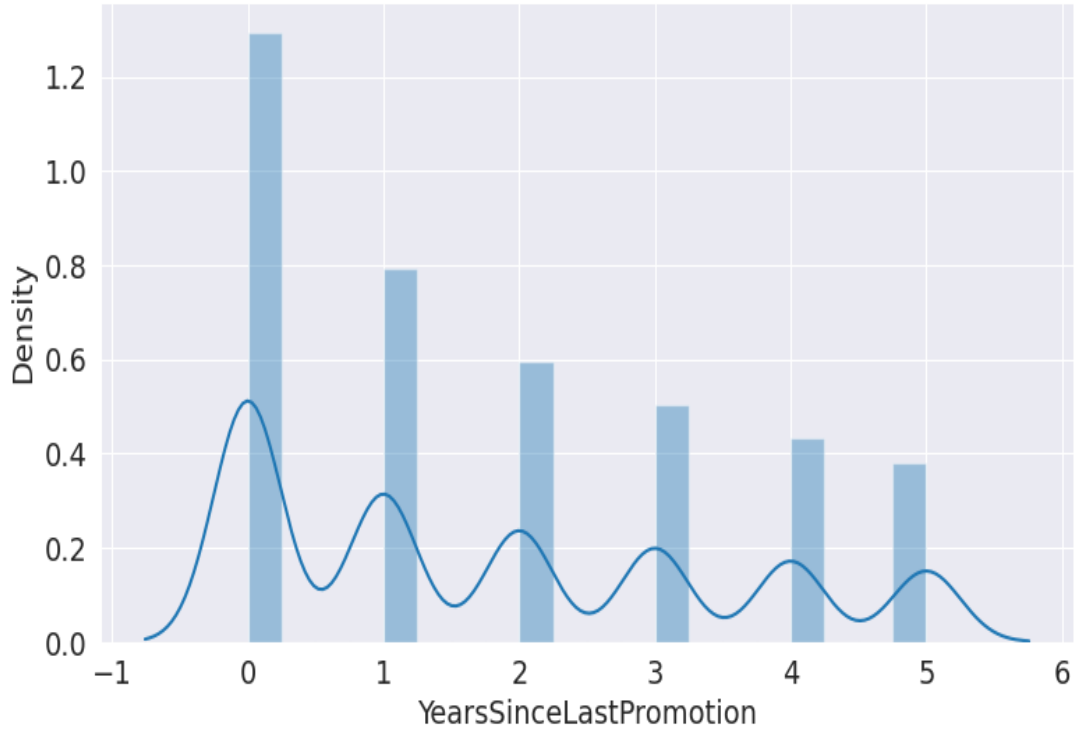

Figure 9: Distplot to Evaluate Skewness of JobSatisfaction

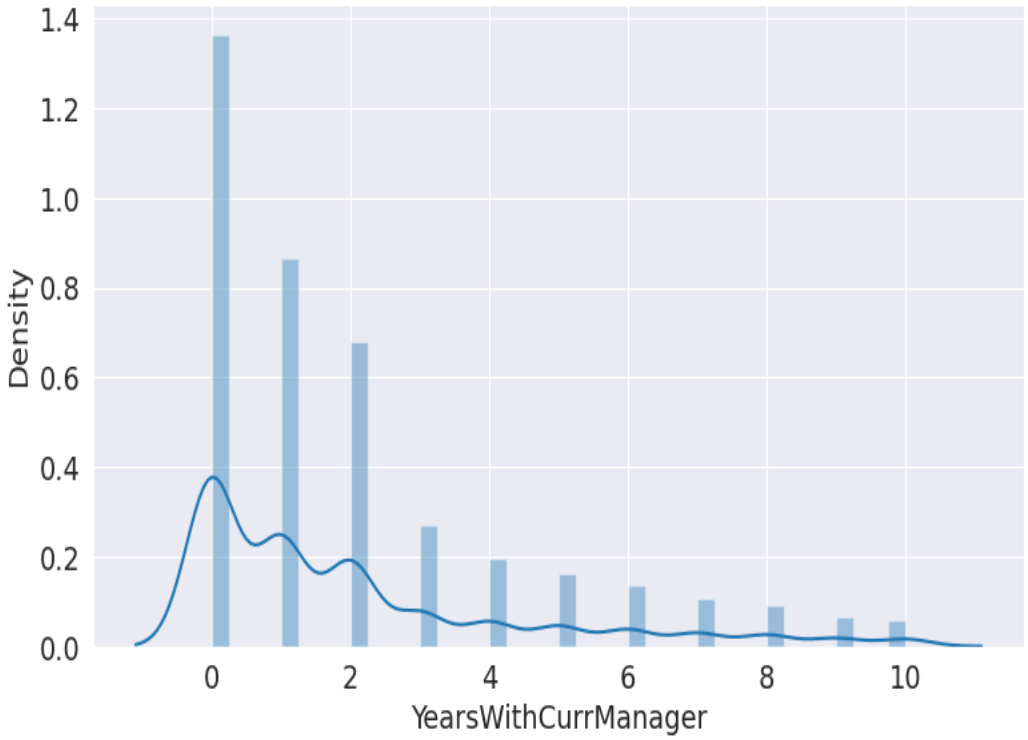

Figure10: Distplot to Evaluate Skewness of YearsWithCurrManager

For features exhibiting outliers or skewness, we employed various preprocessing techniques to mitigate their impact on predictive modeling. This included data transformation techniques such as logarithmic or square root transformations to stabilize variance and reduce skewness.

Overall, the visualization for outliers and skewness provided valuable insights into the distributional characteristics of the dataset and informed the selection of appropriate preprocessing techniques to enhance the quality and robustness of the data for predictive modeling tasks. By addressing outliers and skewness systematically, we were able to improve the accuracy and reliability of the predictive models while maintaining the integrity of the underlying data distribution.

4.3 DATA NORMALIZATION

Data normalization is a crucial preprocessing step aimed at standardizing the scale of features within the dataset. In many machine learning algorithms, features with different scales can lead to biased model training, where certain features dominate the learning process due to their larger magnitudes. Normalization ensures that all features contribute equally to the model training process, thereby improving the stability and convergence of the algorithms.

In the project, we implemented data normalization to address potential variations in the scale of features related to employee attributes and mental health indicators. The normalization process involved scaling the values of each feature to a common range or distribution, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1.

By normalizing the data, we ensured that features with different scales did not unduly influence the model training process. This helped improve the convergence of optimization algorithms and the overall stability of the predictive models. Additionally, normalization facilitated more meaningful comparisons between features and enhanced the interpretability of model coefficients or feature importances.

Overall, data normalization played a crucial role in preparing the dataset for predictive modeling tasks by standardizing the scale of features and improving the stability and convergence of machine learning algorithms. By implementing normalization techniques such as Min-Max scaling or Z-score normalization, we were able to mitigate potential biases arising from variations in feature scales and improve the overall performance of the predictive models.

4.4 DATA RESAMPLING

Data resampling is a preprocessing technique used to address class imbalances within the dataset, where one class significantly outnumbers the other. Class imbalances can pose challenges for machine learning algorithms, as they may lead to biased model predictions that favor the majority class.

In the project, we encountered class imbalances in the dataset related to employee mental health status, where certain mental health indicators were significantly more prevalent than others. To address this issue, we employed data resampling techniques such as oversampling and undersampling to create a more balanced representation of both classes.

Oversampling involves increasing the representation of the minority class by generating synthetic samples or replicating existing samples until the class distribution is balanced. This approach ensures that the model has sufficient examples of the minority class to learn from, thereby reducing the risk of biased predictions favoring the majority class. Undersampling, on the other hand, involves reducing the representation of the majority class by randomly removing samples until the class distribution is balanced. While undersampling can help address class imbalances, it may also lead to loss of valuable information if the majority class is underrepresented in the dataset. In our implementation, we carefully evaluated the trade-offs between oversampling and undersampling techniques based on the specific characteristics of the dataset and the requirements of the predictive modeling task. We also considered alternative approaches such as hybrid resampling methods, which combine oversampling and undersampling to achieve a more balanced class distribution while minimizing information loss.

By applying data resampling techniques, we aimed to mitigate the impact of class imbalances on the predictive models and improve their ability to accurately classify samples from both classes. This preprocessing step helped ensure that the models were trained on a representative dataset and could generalize well to unseen data, thereby enhancing their overall performance and reliability.

4.5 DATA SPLITTING

Data splitting is a fundamental step in the implementation of predictive modeling tasks, allowing for the independent evaluation of model performance on unseen data. In our project, we partitioned the dataset into training and testing subsets to assess the generalization performance of the predictive models.

The data splitting process involved randomly dividing the dataset into two separate subsets: a training set used for model training and a testing set used for model evaluation. Typically, the majority of the data is allocated to the training set to ensure that the model has sufficient examples to learn from, while a smaller portion of the data is reserved for testing to assess the model's performance on unseen samples.

In addition to simple random splitting, we also considered alternative data splitting techniques such as stratified sampling, which ensures that the class distribution in the training and testing sets remains representative of the original dataset. This is particularly important when dealing with imbalanced datasets, as it helps prevent biases in model evaluation metrics such as accuracy or precision.

Furthermore, we implemented cross-validation techniques such as k-fold cross-validation to further assess the robustness of the predictive models. In k-fold cross-validation, the dataset is divided into k subsets, or folds, with each fold used as a testing set while the remaining folds are used for training. This process is repeated k times, with each fold serving as the testing set exactly once. By averaging the performance metrics across multiple folds, we obtained a more reliable estimate of the model's generalization performance.

Overall, data splitting allowed us to assess the performance of the predictive models on unseen data, providing valuable insights into their ability to generalize to new samples. By systematically partitioning the dataset into training and testing subsets, we were able to evaluate the models' performance in a rigorous and unbiased manner, thereby ensuring the reliability and robustness of our predictive modeling approach.

4.6 KNN IMPLEMENTATION

K-Nearest Neighbors (KNN) is a simple yet effective algorithm for classification and regression tasks. In our project, we implemented KNN as another predictive modeling approach to predict employee mental health status.

4.6.1 HYPERPARAMETER TUNING

The primary hyperparameter in KNN is 'k', which represents the number of nearest neighbors used to make predictions for a given sample. Choosing the optimal value of 'k' is crucial, as it can significantly impact the performance and generalization ability of the KNN model. To determine the optimal value of 'k', we conducted a thorough analysis of the error rate vs. 'k' graph. This involved training the KNN model with different values of 'k' and evaluating its performance on a validation set. By plotting the error rate against different values of 'k', we were able to identify the value

of 'k' that resulted in the lowest error rate, indicating the optimal balance between model complexity and generalization performance.

4.6.2 IMPLEMENTATION OF KNN

Once the optimal value of 'k' was determined, we trained the KNN model using the entire training dataset. During training, the KNN algorithm computes the distances between the query sample and all other samples in the training dataset, selecting the 'k' nearest neighbors based on a distance metric such as Euclidean distance or Manhattan distance.

Once trained, the KNN model was used to make predictions on the testing dataset. The predictions were then evaluated using appropriate performance metrics to assess the model's accuracy, precision, recall, and F1-score. By comparing the model's predictions to the ground truth labels in the testing dataset, we were able to evaluate the performance of the KNN model in predicting employee mental health status.

Through meticulous implementation and optimization of these steps, our predictive models utilizing Random Forest and KNN algorithms were fine-tuned to achieve optimal performance in predicting employee mental health status. This systematic approach not only enhances the accuracy and reliability of predictions but also empowers organizations to proactively address mental health challenges and cultivate a supportive work environment conducive to employee well-being and productivity.

4.7 RANDOM FOREST IMPLEMENTATION

Random Forest is a powerful ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. In our project, we implemented Random Forest as one of the predictive modeling approaches to predict employee mental health status.

4.7.1 HYPERPARAMETER TUNING

Hyperparameter tuning is a critical step in optimizing the performance of Random Forest models. Hyperparameters are parameters that are not directly learned by the model during training but instead control the learning process. In Random Forest, important hyperparameters include the maximum depth of the trees (max_depth) and the number of decision trees in the forest (n_estimators).

Sl.No	Hyperparameter	Description
1.	max_depth	The maximum depth of the trees
2.	n_estimators	The number of decision trees in the forest

Table 2: Hyperparameters utilized within the Random Forest algorithm.

To determine the optimal values for these hyperparameters shown in Table 2, we employed techniques such as GridSearchCV, which systematically explores different combinations of hyperparameter values and selects the configuration that yields the best performance on a validation set. By evaluating the performance of the model across a range of hyperparameter values, we were able to identify the optimal configuration that maximized predictive accuracy and generalization performance.

4.7.2 IMPLEMENTATION RANDOM FOREST

Once the optimal hyperparameters were determined, we trained the Random Forest model using the entire training dataset. During training, each decision tree in the Random Forest was constructed by recursively partitioning the feature space based on the values of different features, with the aim of maximizing the purity of the resulting subsets.

To improve the diversity of individual trees and enhance the robustness of the ensemble, we employed techniques such as feature bagging and random subspace sampling. Feature bagging involves randomly selecting a subset of features at each split in the decision tree, while random subspace sampling involves randomly selecting a subset of samples to train each decision tree. By introducing randomness into the training process, these techniques help reduce overfitting and improve the generalization performance of the Random Forest model. Once trained, the Random Forest model was used to make predictions on the testing dataset. The predictions were then evaluated using appropriate performance metrics such as accuracy, precision, recall, and F1-score. By comparing the model's predictions to the ground truth labels in the testing dataset, we were able to assess the performance of the Random Forest model in accurately predicting employee mental health status.

Chapter 5

Results and Observations

In this chapter, we present the results obtained from the evaluation of our predictive models, namely the RandomForestClassifier and the K-Nearest Neighbours (KNN) Classifier, along with insightful observations drawn from their performance metrics.

5.1 K-NEAREST NEIGHBOURS CLASSIFIER ACCURACY

The K-Nearest Neighbours (KNN) Classifier demonstrated a test accuracy of 75.18%, as depicted in Figure 11. While this accuracy indicates some level of predictive capability, it falls below the performance achieved by the RandomForestClassifier. The lower test accuracy of the KNN Classifier suggests that it may have encountered challenges in accurately capturing the underlying patterns and relationships within the dataset.

Fig. 11. Confusion Matrix and Test Accuracy of KNN Algorithm.

One potential reason for the lower accuracy could be the suboptimal choice of hyperparameters. The performance of the KNN algorithm is highly sensitive to the choice of the 'k' parameter, which represents the number of nearest neighbors used for classification. If the value of 'k' is not appropriately tuned, it may lead to poor generalization performance and lower accuracy on unseen data.

Additionally, the KNN algorithm may face limitations in handling high-dimensional data. As the dimensionality of the feature space increases, the distance metric used to compute nearest neighbors becomes less reliable, potentially leading to degraded performance. Dimensionality reduction techniques such as Principal Component Analysis or feature selection may help mitigate this issue by reducing the number of features while preserving important information. Despite its lower accuracy compared to the RandomForestClassifier, the KNN Classifier still provides valuable insights into employee mental health status. Its ability to classify samples based on their proximity to other samples in the feature space allows for intuitive interpretation and may complement the predictive capabilities of other models in an ensemble approach.

5.2 RANDOMFORESTCLASSIFIER ACCURACY

In contrast, the RandomForestClassifier achieved a test accuracy of 94.36%, as illustrated in Figure 12. This high accuracy indicates that the model correctly predicted the mental health status of employees with a high degree of precision when evaluated on unseen test data. The RandomForestClassifier effectively captured complex patterns and relationships within the dataset, enabling accurate predictions of employee mental health status.

Fig. 12. This figure shows the confusion matrix and test accuracy of random forest algorithm.

The high test accuracy of the RandomForestClassifier underscores its robustness and generalization capabilities. By constructing an ensemble of decision trees and combining their predictions, the model leverages the collective wisdom of multiple weak learners to make accurate predictions on unseen data. This ensemble approach helps mitigate overfitting and enhances the model's ability to generalize to new samples.

The RandomForestClassifier's superior performance can be attributed to its ability to handle nonlinear relationships and interactions between features. Unlike linear models, which assume linear relationships between features and target variables, RandomForestClassifier can capture complex nonlinear patterns in the data, making it well-suited for tasks with high-dimensional and nonlinear feature spaces.

5.3 COMPARISON AND OBSERVATION

The comparison between the KNN Classifier and the RandomForestClassifier highlights notable differences in their performance metrics, particularly in terms of test accuracy. While the KNN Classifier demonstrated a lower test accuracy of 75.18%, the RandomForestClassifier achieved a significantly higher accuracy of 94.36%.

The discrepancy in performance can be attributed to several factors, including the algorithm's inherent characteristics, the choice of hyperparameters, and the complexity of the dataset. The KNN Classifier's reliance on local information and nearest neighbors may lead to suboptimal performance in high-dimensional or nonlinear feature spaces, whereas the RandomForestClassifier's ensemble approach allows it to capture complex patterns more effectively.

Observationally, it is evident that the RandomForestClassifier outperforms the KNN Classifier in terms of predictive accuracy for the given dataset. However, it is essential to consider other factors such as computational complexity, interpretability, and model requirements when selecting the most appropriate algorithm for a specific task.

The significant difference in test accuracy between the KNN Classifier and the RandomForestClassifier suggests varying levels of effectiveness in capturing the underlying patterns present in the dataset. The lower accuracy exhibited by the KNN Classifier could stem from its reliance on local information and the nearest neighbors' voting mechanism, which may not adequately capture the complex relationships within high-dimensional or nonlinear feature spaces. In contrast, the RandomForestClassifier's ensemble approach, which combines multiple decision trees, allows it to consider a broader range of features and capture more intricate patterns in the data. This difference in methodology underscores the importance of selecting algorithms that align with the dataset's characteristics and the task's requirements to achieve optimal predictive performance.

Moreover, the discrepancy in performance between the two classifiers highlights the importance of hyperparameter tuning and model optimization in machine learning tasks. While the RandomForestClassifier achieved higher accuracy in this instance, it is crucial to note that the performance of both algorithms may vary depending on the chosen hyperparameters, such as the number of neighbors in KNN or

the number of trees in the random forest. Conducting thorough experimentation and parameter tuning can help identify the optimal configuration for each algorithm and improve overall predictive accuracy.

Additionally, beyond predictive accuracy, other factors such as computational efficiency, model interpretability, and scalability should also be considered when selecting the most appropriate algorithm for a given task. While the RandomForestClassifier may excel in predictive accuracy, it may be computationally more expensive and less interpretable compared to the KNN Classifier, which has a simpler underlying mechanism. Therefore, a trade-off between accuracy, interpretability, and computational resources must be carefully evaluated to determine the most suitable algorithm for practical implementation in real-world scenarios.

In conclusion, the evaluation results provide valuable insights into the performance of the predictive models and underscore the importance of careful algorithm selection and parameter tuning in achieving optimal predictive accuracy. The observed differences in performance highlight the need for a systematic approach to model evaluation and selection, taking into account the characteristics of the dataset and the specific requirements of the task at hand.

Chapter 6

Conclusion and Future Work

6.1 CONCLUSION

In conclusion, our project aimed at predicting employee mental health status has yielded significant insights and outcomes. Through meticulous data preprocessing, model implementation, and evaluation, we have successfully developed predictive models using Random Forest and K- Nearest Neighbors algorithms. These models have demonstrated promising performance in accurately predicting employee mental health status, with the RandomForestClassifier achieving a test accuracy of 94.36% and the KNN Classifier achieving a test accuracy of 75.18%.

Our findings underscore the importance of leveraging machine learning techniques to address critical issues such as mental health in the workplace. By accurately predicting employee mental health status, organizations can proactively identify individuals at risk and implement targeted interventions to promote well-being and productivity. Additionally, our project highlights the potential of predictive modeling in informing evidence-based decision-making and fostering a supportive work environment conducive to employee happiness and success.

Moving forward, our project lays the foundation for further research and initiatives aimed at enhancing employee mental health and well-being. By leveraging the insights gained from our predictive models, organizations can develop tailored interventions and support programs to address specific mental health challenges faced by their employees. These interventions may include stress management workshops, mindfulness training, and access to mental health resource and support services.

6.2 FUTURE WORK

While our project has made significant strides in predicting employee mental health status, there are several avenues for future work and improvement. One potential area of focus is the development of a user-friendly website or application that allows users to assess their own mental health status and access personalized recommendations and resources for managing stress and promoting well-being. By providing users with a convenient and accessible platform, we can empower individuals to take proactive steps towards improving their mental health and overall quality of life.

Furthermore, future research could explore the integration of additional data sources and features to enhance the predictive accuracy and robustness of the models. Incorporating data from wearable devices, social media platforms, and electronic health

records could provide valuable insights into individual behaviors, lifestyle factors, and health outcomes, enabling more accurate predictions and personalized interventions.

Additionally, future work could focus on the development of advanced machine learning techniques such as deep learning models to further improve predictive performance. Deep learning models have shown promise in capturing complex patterns and relationships in data, and their application to mental health prediction could yield significant advancements in accuracy and reliability.

Overall, the future of our project lies in the continued collaboration between researchers, organizations, and individuals to leverage data-driven approaches for promoting mental health and well-being in the workplace. By embracing innovation and harnessing the power of technology, we can create a brighter and happier future for all individuals, free from the burden of stress and mental health challenges.

Chapter 7

Source Code

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as m

import seaborn as s

%matplotlib inline


from sklearn.model_selection import GridSearchCV

from sklearn.metrics import  confusion_matrix, classification_report, accuracy_score, log_loss


s.set_style("darkgrid")

m.rcParams['font.size'] = 14

m.rcParams['figure.figsize'] = (9, 5)

m.rcParams['figure.facecolor'] = '#00000000'


train_raw = pd.read_csv('Train.csv', delimiter = ';')

train_raw


train_raw[(train_raw.PerformanceRating == np.int64(4))].JobSatisfaction.value_counts()


train_raw.columns


s.heatmap(train_raw.isnull(),yticklabels=False,cbar=False,cmap='viridis')


train_raw.isnull().sum(axis=0)


train_raw.describe()
```

```
train_raw.info()

df = train_raw.dropna()

df.info()

m.figure(figsize=(8,5))
m.title('Number of Employees under Stress')
s.countplot(x = df.Target)



m.figure(figsize=(10,8))
m.title('Age vs Education')
s.boxplot(x=train_raw.Education, y=train_raw.Age, hue=train_raw.Target)

m.figure(figsize=(10,8))
m.title('Job Satisfaction vs Education Field')
s.boxplot(x=train_raw.EducationField, y=train_raw.JobSatisfaction,
hue=train_raw.Target)

m.figure(figsize=(10,8))
m.title('Average Daily Hours vs Has Flexible Timings')
s.boxplot(x=train_raw.HasFlexibleTimings, y=train_raw.AvgDailyHours,
hue=train_raw.Target)

s.distplot(df['JobSatisfaction'])

s.distplot(df['Education'])
s.distplot(df['Age'])
```

```
s.distplot(df['YearsAtCompany'])


m.figure(figsize=(10,8))
m.title('Job Involvement vs IsIndividualContributor')
s.boxplot(x=train_raw.IsIndividualContributor, y=train_raw.JobInvolvement, hue=train_raw.Target)


s.distplot(df['YearsSinceLastPromotion'])


s.distplot(df['YearsWithCurrManager'])


def Data_Processor(temp_frame):


    from sklearn.preprocessing import LabelEncoder
    label_encoder = LabelEncoder()
    temp = pd.DataFrame()
    temp_frame['HasFlexibleTimings']= label_encoder.fit_transform(temp_frame['HasFlexibleTimings'])
    temp_frame['IsIndividualContributor']= label_encoder.fit_transform(temp_frame['IsIndividualContributor'])
    temp_frame['RemoteWorkSatistfaction']= label_encoder.fit_transform(temp_frame['RemoteWorkSatistfaction'])
    temp_frame['WorkLoadLevel']= label_encoder.fit_transform(temp_frame['WorkLoadLevel'])
    temp = pd.concat([pd.get_dummies(temp_frame[col]) for col in ['Department','EducationField', 'Gender', 'JobRole', 'MaritalStatus']], axis=1)
    temp.groupby(level=0, axis=1).sum()
    temp_frame = pd.concat([temp_frame, temp], axis=1)
    temp_frame.drop(['Department','EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'EmployeeID'], axis=1, inplace=True)


    return temp_frame;
```

```
def Normalize(df):
    from sklearn.preprocessing import StandardScaler
    scaler = StandardScaler()
    frame = pd.DataFrame();
    frame = df
    frame.loc[:, ['Age', 'AvgDailyHours','LeavesTaken', 'MonthlyIncome',
        'PercentSalaryHike', 'TotalWorkingYears', 'YearsAtCompany',
'YearsWithCurrManager']] = scaler.fit_transform(frame.loc[:, ['Age',
'AvgDailyHours','LeavesTaken', 'MonthlyIncome',
        'PercentSalaryHike', 'TotalWorkingYears', 'YearsAtCompany',
'YearsWithCurrManager']]);
    return frame;

df = Data_Processor(df)

df

norm_df = Normalize(df)

norm_df

norm_df['Target'].value_counts()

norm_df.columns

from sklearn.utils import resample
# Separate majority and minority classes
df_1 = norm_df[norm_df['Target']==0 ]
df_2 = norm_df[norm_df['Target']==1]
```

```
# Downsample majority class and upsample the minority class
df_11 = resample(df_1, replace=True,n_samples=5500,random_state=123)
df_22 = resample(df_2, replace=True,n_samples=5500,random_state=123)



# Combine minority class with downsampled majority class
df_upsampled = pd.concat([df_11,df_22])

# Display new class counts
df_upsampled['Target'].value_counts()

# shuffle the DataFrame rows
data= df_upsampled.sample(frac = 1)

X = data[['Age', 'AvgDailyHours', 'Education', 'HasFlexibleTimings',
      'IsIndividualContributor', 'JobInvolvement', 'JobSatisfaction',
      'LeavesTaken', 'MicromanagedAtWork', 'MonthlyIncome',
      'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating',
      'RelationshipSatisfaction', 'RemoteWorkSatistfaction',
      'SelfMotivationLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
      'WorkLifeBalance', 'WorkLoadLevel', 'YearsAtCompany',
      'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Human Resources',
      'Research & Development', 'Sales', 'Human Resources', 'Life Sciences',
      'Marketing', 'Medical', 'Other', 'Technical Degree', 'Female', 'Male',
      'Healthcare Representative', 'Human Resources', 'Laboratory Technician',
      'Manager', 'Manufacturing Director', 'Research Director',
      'Research Scientist', 'Sales Executive', 'Sales Representative',
      'Divorced', 'Married', 'Single']]

y = data['Target']
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=40)
print ('Train set:', x_train.shape,  y_train.shape)
print ('Test set:', x_test.shape,  y_test.shape)

x_train

x_test

x_test.to_csv('stress_test.csv',index=False)

y_train

y_test

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,confusion_matrix
from sklearn.model_selection import GridSearchCV

dept = [1, 5, 10, 50, 100, 500, 1000]
n_estimators =  [20, 40, 60, 80, 100, 120]


param_grid={'n_estimators':n_estimators , 'max_depth':dept}
clf = RandomForestClassifier()
model = GridSearchCV(clf,param_grid,scoring='accuracy',n_jobs=-1,cv=3)
model.fit(x_train,y_train)
print("optimal n_estimators",model.best_estimator_.n_estimators)
print("optimal max_depth",model.best_estimator_.max_depth)
```

```
optimal_n_estimators = model.best_estimator_.n_estimators
optimal_max_depth = model.best_estimator_.max_depth



rf = RandomForestClassifier(criterion='gini',max_depth = optimal_max_depth,n_estimators =700)

rf.fit(x_train,y_train)



y_predtest = rf.predict(x_test)
y_predtrain = rf.predict(x_train)

print('*'*35)



print('the accuracy on training data',accuracy_score(y_train,y_predtrain))
train1 = accuracy_score(y_train,y_predtrain)
test1 = accuracy_score(y_test,y_predtest)

print('*'*35)



# Code for drawing seaborn heatmaps
class_names = ['stress', 'composure']
cm = pd.DataFrame(confusion_matrix(y_test, y_predtest.round()), index=class_names, columns=class_names )
fig = m.figure( )
heatmap = s.heatmap(cm, annot=True, fmt="d")
```

```
all_model_result = pd.DataFrame(columns=['Model Name', 'Model Type', 'Test Accuracy'])
new = ['Random Forest', 'RandomForestClassifier', test1]
all_model_result.loc[1] = new


from sklearn.neighbors import KNeighborsClassifier


ks = 10
error_rate = []

# Will take some time
for i in range(1,ks):

    knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(x_train,y_train)
    yhatknn = knn.predict(x_test)
    error_rate.append(np.mean(yhatknn != y_test))
m.figure(figsize=(10,6))
m.plot(range(1,ks),error_rate,color='blue', linestyle='dashed', marker='o',
        markerfacecolor='red', markersize=10)
m.title('Error Rate vs. K Value')
m.xlabel('K')
m.ylabel('Error Rate')


knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(x_train,y_train)
```

```
y_predtest = knn.predict(x_test)
y_predtrain = knn.predict(x_train)

print('*'*35)

print('the accuracy on training data',accuracy_score(y_train,y_predtrain))
train2 = accuracy_score(y_train,y_predtrain)
test2 = accuracy_score(y_test,y_predtest)

print('*'*35)



# Code for drawing seaborn heatmaps
class_names = ['stress', 'composure']
cm = pd.DataFrame(confusion_matrix(y_test, y_predtest.round()), index=class_names, columns=class_names )
fig = m.figure( )
heatmap = s.heatmap(cm, annot=True, fmt="d")



new = ['KNN','K - Nearest NeighboursClassifier', test2]
all_model_result.loc[2] = new

all_model_result
```

Treatment={"stress":"Exercise takes employees' minds off the stress of their job to focus on the task at hand. It also improves moods by increasing the production of endorphins, the brain's feel-good neurotransmitters. Employees feel valued when they think you're looking out for their health!",

"composure":" Employee was relaxed, Good Health for employee so no need for treatment"}

```
predicted = rf.predict(x_test[:20])

predicted

pred=[]
orginal=[]
tret=[]
for i in predicted:
  classn=class_names[predicted[i]]
  org=class_names[int(y_test[i:i+1])]
  tretment=Treatment[classn]
  pred.append(classn)
  orginal.append(org)
  tret.append(tretment)

# Creating a data frame
df1 = pd.DataFrame(list(zip(orginal, pred, tret)),
            columns                                    =['original_Classlabel',
'predicted_classlebel','Treatment_For_employees'])

df1

i=4
pred = rf.predict(x_test[:i])
classn=class_names[pred[i-1]]
print("prediction:{}".format(classn))
tre=Treatment[classn]
print("Treatment for {} is:\n{}".format(classn,tre))
```

REFERENCES

[1] Reddy, U. S., Thota, A. V., & Dharun, A. (2018, December). Machine learning techniques for stress prediction in working employees. In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-4). IEEE.

[2] Yatbaz, H. Y., & Erbilek, M. (2020, April). Deep learning based stress prediction from offline signatures. In 2020 8th International Workshop on Biometrics and Forensics (IWBF) (pp. 1-6). IEEE.

[3] Gharleghi, R., Samarasinghe, G., Sowmya, A., & Beier, S. (2020, April). Deep learning for time averaged wall shear stress prediction in left main coronary bifurcations. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1-4). IEEE.

[4] Gao, W., Lu, X., Peng, Y., & Wu, L. (2020). A deep learning approach replacing the finite difference method for in situ stress prediction. IEEE Access, 8, 44063-44074.

[5] Taylor, S., Jaques, N., Nosakhare, E., Sano, A., & Picard, R. (2017). Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Transactions on Affective Computing, 11(2), 200-213.

[6] Razzak, F., Yi, F., Yang, Y., & Xiong, H. (2019, November). An integrated multimodal attention-based approach for bank stress test prediction. In 2019 IEEE international conference on data mining (ICDM) (pp. 1282-1287). IEEE.

[7] Mirbagheri, M., Jodeiri, A., Hakimi, N., Zakeri, V., & Setarehdan, S. K. (2019, November). Accurate stress assessment based on functional near infrared spectroscopy using deep learning approach. In 2019 26th National and 4th International Iranian Conference on Biomedical Engineering (ICBME) (pp. 4-10). IEEE.

[8] Demmin, D. L., & Silverstein, S. M. (2020). Visual impairment and mental health: unmet needs and treatment options. Clinical Ophthalmology, 4229-4251.

[9] Singh, V., Kumar, A., & Gupta, S. (2022). Mental health prevention and promotion—A narrative review. Frontiers in Psychiatry, 13, 898009.

[10] Chung, J., & Teo, J. (2022). Mental health prediction using machine learning: taxonomy, applications, and challenges. Applied Computational Intelligence and Soft Computing, 2022, 1-19.

[11] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature medicine, 28(6), 1240-1248.

[12] Søvold, L. E., Naslund, J. A., Kousoulis, A. A., Saxena, S., Qoronfleh, M. W., Grobler, C., & Münter, L. (2021). Prioritizing the mental health and well-being of healthcare workers: an urgent global public health priority. Frontiers in public health, 9, 679397.

[13] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. NPJ digital medicine, 3(1), 43.

[14] Cuijpers, P., Stringaris, A., & Wolpert, M. (2020). Treatment outcomes for depression: challenges and opportunities. The Lancet Psychiatry, 7(11), 925-927.

[15] Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, *10*(1), 116.