Taltech - Tallin University of Technology
Business Analytics Course - MMA3100
October 27, 2024

Mental Health in Tech Companies

A Data Driven Approach to Mental Health Diseases

Project Report

Bozhok Yaroslava
email: yabozh@taltech.ee

Makarov Valeri
email: vamaka@taltech.ee

Navarra Filippo
email: finava@taltech.ee

Course Related Group Project

1 Introduction

Mental health in the workplace is an increasingly important issue, particularly in fast-paced and high-stress industries like technology. In recent years, there has been growing awareness around mental health disorders and their impact on employee productivity, well-being, and overall organizational success. Addressing this issue is critical for companies striving to create supportive, productive, and healthy work environments. To this end, the Open Sourcing Mental Illness (OSMI) Mental Health in Tech Survey 2016 dataset provides valuable insights into the mental health conditions of employees in the tech sector.

This report focuses on exploring the mental health landscape of employees in the tech industry by analyzing the 2016 OSMI dataset. The aim of the analysis is to identify patterns and relationships in the data that can help organizations better understand mental health concerns in the workplace. Specifically, the report will address the business problem of how mental health awareness, resources, and workplace culture impact employees'.

Through the use of data preprocessing, visualizations, and machine learning techniques, the report will provide key insights into the factors influencing mental health in the workplace. These insights will help tech companies, and mainly HR department, develop better mental health policies and offer more targeted resources to employees, ultimately creating a healthier and more productive work environment.

Key steps of this project include:

- **Data Cleaning**: Checking for missing values and removing redundant information.
- **Data Visualization**: Visualizing trends and distributions related to mental health issues and workplace conditions.
- **Modeling**: Using machine learning models to predict mental health outcomes based on various workplace factors.
- **Conclusions**: Offering recommendations to improve workplace mental health support based on data-driven insights.

The analysis performed here aligns with the core objective of understanding and improving mental health in the workplace, which is essential for both employee well-being and organizational success.

2 Data Cleaning

The data cleaning process is a critical step in ensuring the dataset's quality before performing any analysis or modeling. The OSMI Mental Health in Tech Survey 2016 dataset initially contained 1433 rows and 63 columns, which covered a wide variety of topics related to mental health in the workplace. However, upon inspection, we found significant issues with missing values, redundant variables, and inconsistent data formats. Therefore, we implemented a structured data cleaning approach to address these issues and prepare the dataset for further analysis.

2.1 Handling Missing Values

The first task was to address missing values, which are common in survey datasets. Our process has seen several steps.

2.1.1 Dropping Columns with High Missingness

We decided to drop columns where more than 50% of the data was missing, as these columns would likely introduce too much noise into the analysis. The following columns were dropped: *'Is your primary role within your company related to tech/IT?' 'Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?' 'Do you know local or online resources to seek help for a mental health disorder?' "If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?" "If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?" "If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?" "If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?" "Do you believe your productivity is ever affected by a mental health issue?" "If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?" "Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?" "Do you know the options for mental health care available under your employer-provided coverage?"*

After this step, the dataset was reduced to 1433 rows × 52 columns.

2.1.2 Handling Repeated Missing Values

Some variables had missing values that were highly correlated. For instance, many self-employed workers did not have responses for workplace-related questions. We handled this by removing rows with 287 missing values for *"Does your employer provide mental health benefits as part of healthcare coverage?"*, reducing the dataset to 1015 rows × 52 columns. Eliminating missing values for this variable, allowed us to get rid of other NAN in workplace-related questions. Specifically for variables: *"How many employees does your company or* *organization have?", "Is your employer primarily a tech company/organization?", "Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?", "Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?", "Does your employer offer resources to learn more about mental health concerns and options for seeking help?", "If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:", "Do you think that discussing a mental health disorder with your employer would have negative consequences?", "Do you think that discussing a physical health issue with your employer would have negative consequences?", "Would you feel comfortable discussing a mental health disorder with your coworkers?", "Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?", "Do you feel that your employer takes mental health as seriously as physical health?", "Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?".*

Another step has seen dropping missing values for the specific column *"Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?"*, due the difficult imputation of the nature of NAN.

The dataset shape is now 976 x 52.

2.1.3 Handling Conditional Missing Values

Certain questions were conditional upon a previous response. For example, respondents were only asked about diagnosed mental health conditions if they had previously indicated that they had a mental health disorder. To manage these conditional responses, we replaced missing values with "Not Applicable" in the following columns:

"What US state or territory do you work in?" "What US state or territory do you live in?" "Why or why not?" "If yes, what condition(s) have you been diagnosed with?" "If maybe, what condition(s) do you believe you have?" "If so, what condition(s) were you diagnosed with?"

2.1.4 Standardizing Categorical Variables

The gender variable presented several challenges, as the dataset included a variety of terms for gender, including 'Male', 'male', 'Male ', 'Female', 'F', 'Cis female', and many others. To simplify this variable and ensure consistency, we grouped the responses into three categories: Male, Female, and Other. This categorization allowed for easier analysis while still respecting the diversity of gender identities present in the dataset.

2.2 Redundant Columns

The column *"Are you self-employed?"* was deemed irrelevant to the core purpose of the analysis (which focuses on employees' experiences within organizational settings) and was thus removed. After removing this, the dataset had 976 rows × 51 columns.

The column *"Do you have previous employers?"* was removed because it contained only a single unique value (all responses were 1), meaning it would not provide any useful information for analysis.

After completing these steps, the final cleaned dataset had 976 rows and 50 columns, making it a more manageable and reliable dataset for analysis.

2.3 Final Data Check

To ensure the data was ready for analysis, we conducted the following checks:

- Data Types: Ensured that each variable was in the appropriate data type (e.g., categorical variables were encoded as object or category, numeric variables were integers or floats).
- Missing Values: After the above cleaning steps, no significant missing values remained in the dataset.
- Outlier Detection: We briefly checked for any extreme outliers that could skew results, but no severe issues were found at this stage. The only change we made was related to an "What is your age" observation were the age 323, 3 and 99 years old was detected and replaced with 32 for the first and the most frequent value *30* for the latter two values.

Through careful data cleaning, we were able to transform a dataset with significant missing data and inconsistencies into a clean and manageable dataset. The resulting dataset, which now consists of 976 rows × 50 columns, is ready for analysis and modeling, with all key issues around missing data and variable standardization addressed. This process ensures the integrity and quality of the subsequent analysis and modeling steps.

3 Data Visualization

After the data cleaning process, we moved to the visualization step, where we explored various aspects of the dataset to gain insights into the mental health landscape in the tech workplace. Visualizing data helps to uncover patterns, relationships, and trends that might not be immediately apparent from the raw data.

The key objectives of the visualization process were:

- To understand the distribution of important variables in the dataset.
- To identify common mental health disorders in the tech workplace.
- To explore demographic distributions related to age, gender, and company size.

To start, we used a range of histograms, bar plots and maps, to explore the distribution of key categorical and numeric variables. This provided us with a basic understanding of the data structure and helped in identifying any potential biases or interesting trends.

Company Size: We plotted a histogram of *"How many employees does your company or organization have?"*(Figure1) to understand the distribution of company sizes. This helped us identify whether most respondents worked in small, medium, or large companies, and whether this factor influenced mental health outcomes.

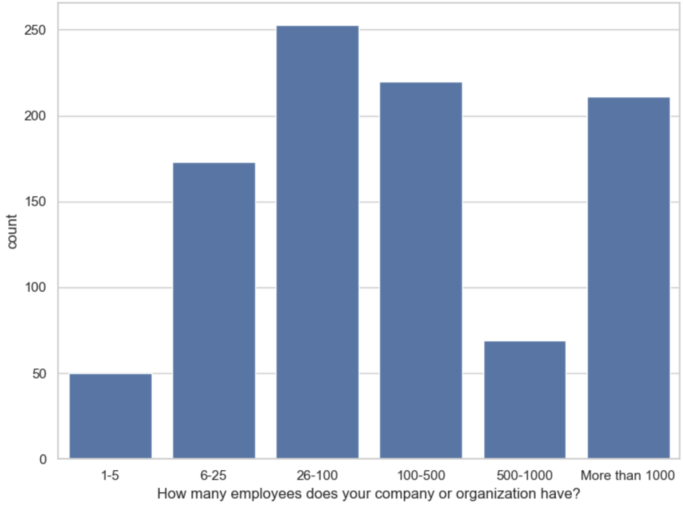


Figure 1: Company Size Respondents

Mental Health Resources: We decided to visualize some variables related to the possibilities companies offer to their employers, such as *"Has your employer ever formally discussed mental health?"*(Figure2) and *"Did your previous employers provide resources to learn more about mental health issues and how to seek help?"*(Figure3) were visualized to assess how frequently these discussions and resources are available in tech companies.

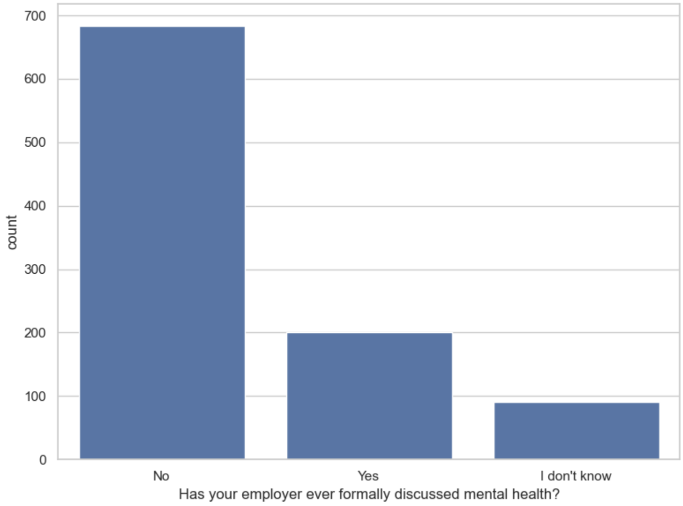


Figure 2: Mental Health Formal Discussion

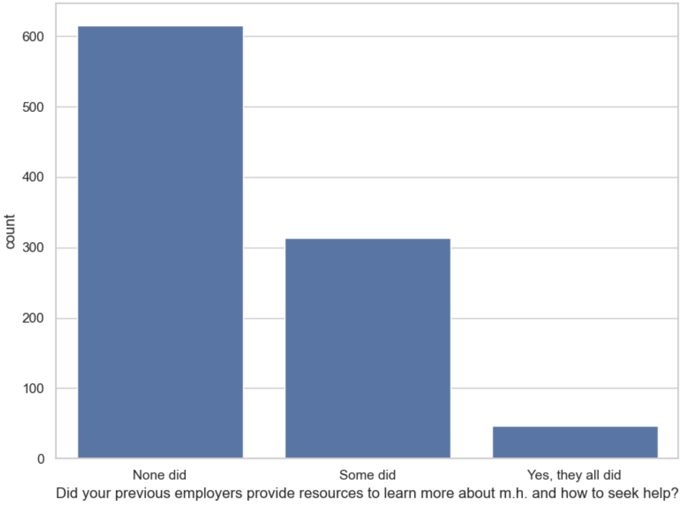


Figure 3: Previous Employers Resources

Comfort in Discussing Mental Health: The questions *"Would you bring up a mental health issue with a potential employer in an interview?*(Figure4) and *"Do you think that discussing a mental health disorder with previous employers would have negative consequences?"*(Figure5) were also visualized, helping to gauge the openness of mental health discussions within organizations.

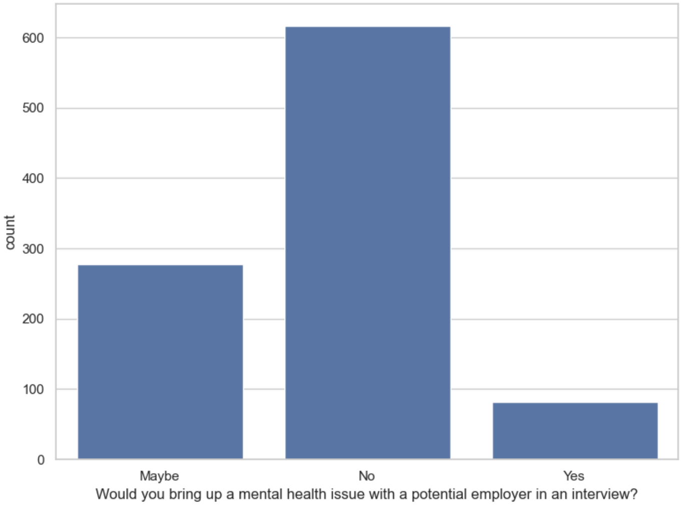


Figure 4: Bring up MHD in an interview

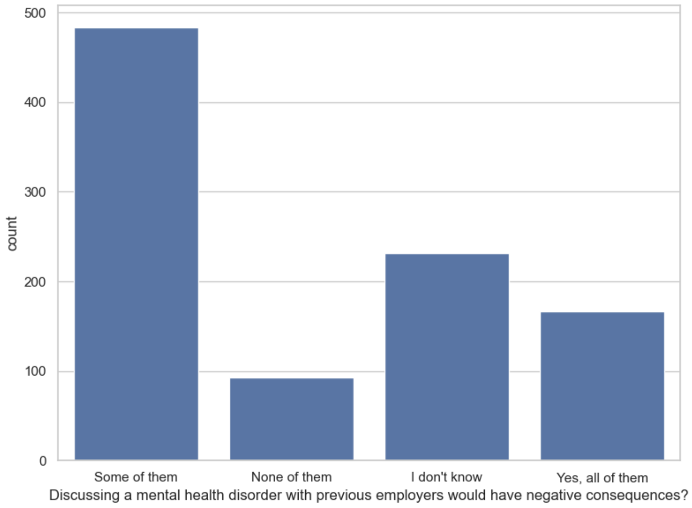


Figure 5: Discuss MHD with previous employers

Personal Information: To better understand who our respondents are, we decided to plot personal information such as *"What is your age"*(Figure6), *"What is your gender"*(Figure7) and *"What country do you live in?"*(Figure8).

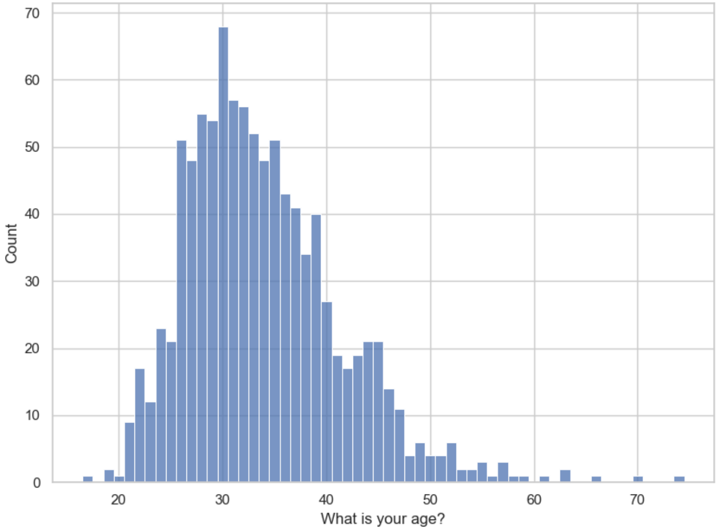


Figure 6: Age Distribution

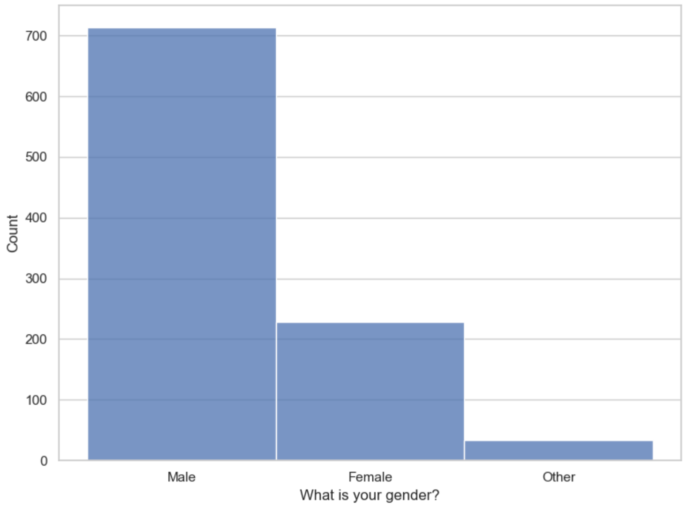


Figure 7: What is your gender

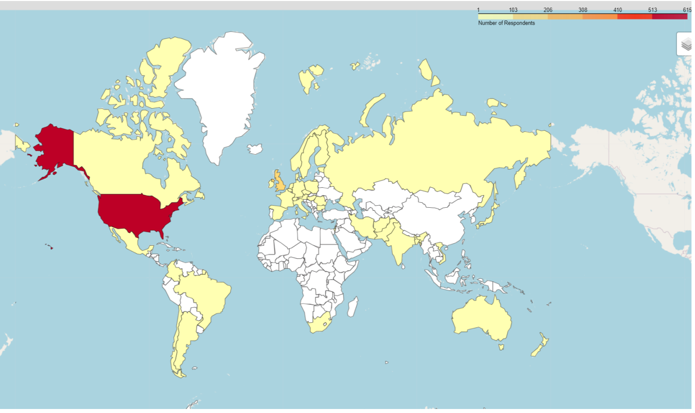


Figure 8: Where our respondents are from

The majority of our survey respondents fall within the 25 to 40 age range, with a significant portion identifying as male. Geographically, the respondents are predominantly based in the United States, reflecting the demographics of the tech industry. This age range is particularly relevant, as individuals in this group are often in the midst of their careers, which makes understanding their mental health challenges within the workplace especially critical.

We considered it important to examine whether employees had family members who previously suffered from mental health conditions (Figure9). This factor could potentially influence both their perception of mental health issues and their sensitivity toward the topic. Individuals with a family history of mental health disorders may be more aware of mental health challenges, which could impact their attitudes, openness to seeking help, and how they engage with workplace mental health resources.

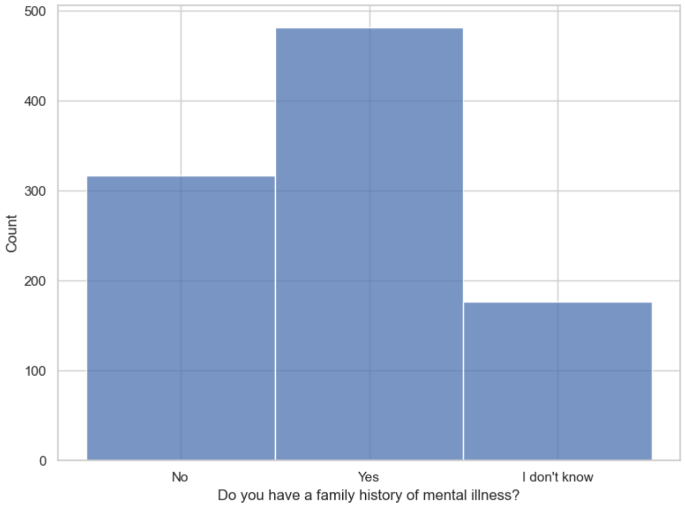


Figure 9: History of Mental Illnesses

Mental Health Disorder Distribution: We visualized the distribution of common mental health disorders using bar plots (Figure10), focusing on diagnoses like ”Anxiety Disorder” ”ADHD”, ”OCD” and ”PTSD”. Since respondents could select multiple options, we assumed their first choice was the most significant and grouped the disorders accordingly.

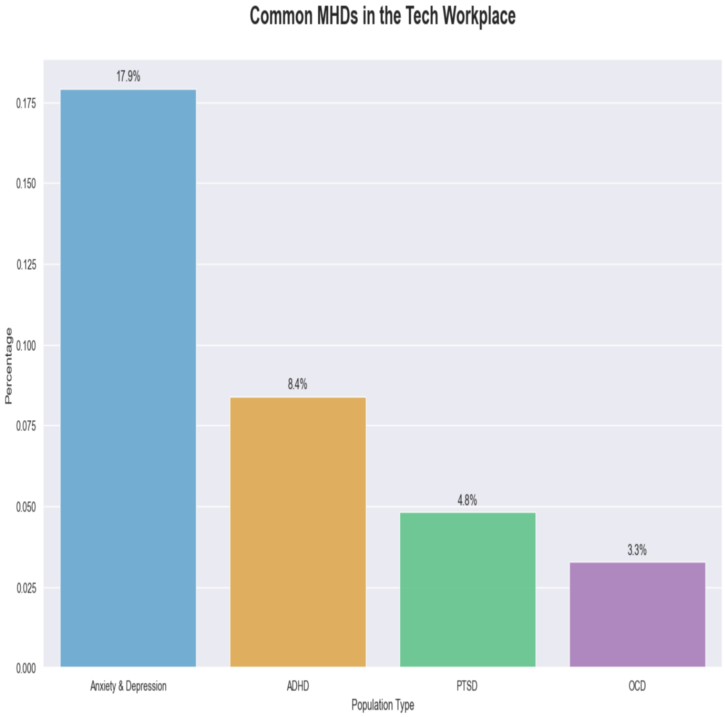


Figure 10: Most common MHD

We also explored the gender distribution of diagnosed mental health disorders using stacked bar plots (Figure11) to see how mental health conditions varied between different genders (Male, Female, and Other).

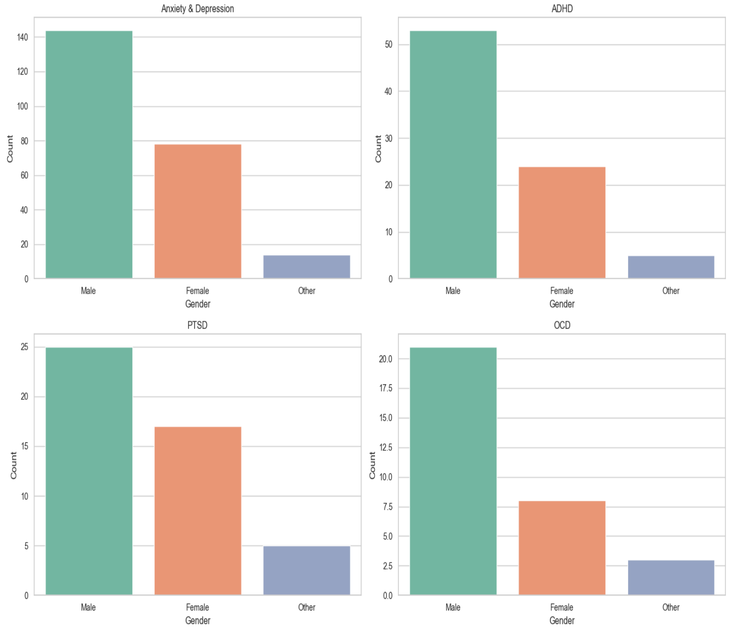


Figure 11: Most common MHD per Gender

Mental Health Disorders and Demographics Boxplot was used to examine how most common disorders are distributed across different age groups (Figure12).

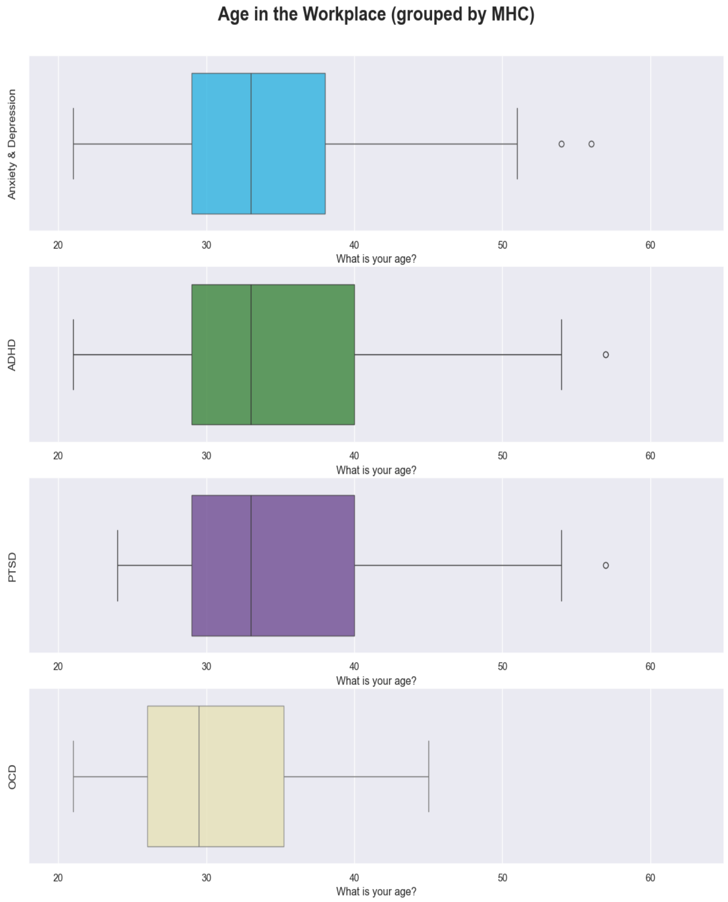


Figure 12: BoxPlot Age and MHD

The age distribution across mental health conditions tends to center around the 25-40 age group, which aligns with the typical working age for many professionals in the tech industry. Outliers in each category are rare, meaning most mental health issues are concentrated within a specific age range rather than being spread across all ages.

A map of most common diseases per country has been plotted (Figure 13). For this map we decided to use the more general annotations of Anxiety Disorder and Mood Disorder. Anxiety is the most widespread, affecting many countries in Europe, Canada, Russia, and Asia. Mood disorders are prevalent in nations like the United States, United Kingdom, Chile and Australia. ADHD is the most common disease in Denmark.

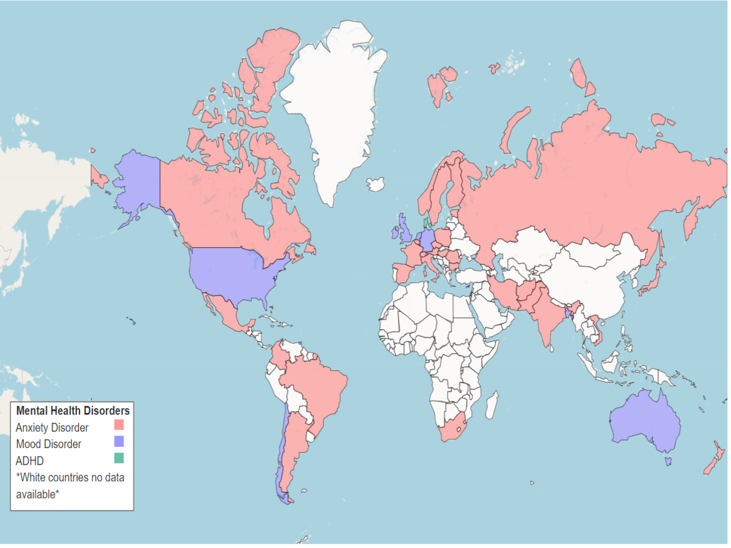
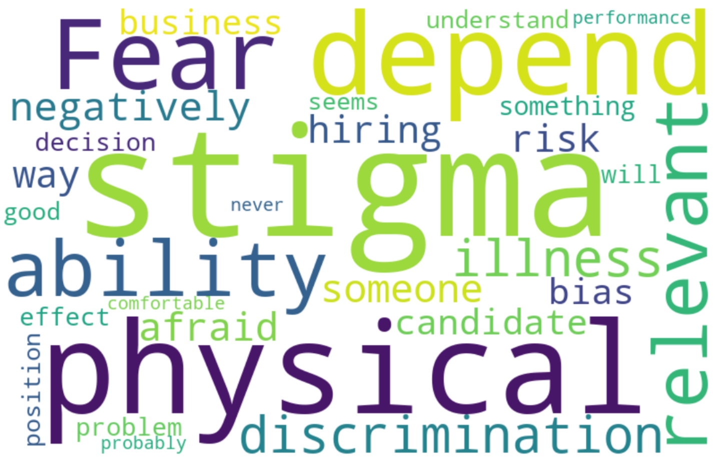


Figure 13: Most Common MHD per Country

Word Cloud Visualization: This word cloud visually represents the most common words employees used when explaining why or why not they would bring up a mental health disorder (MHD) with a potential employer during an interview. The size of each word correlates with its frequency in the responses, giving a quick snapshot of key concerns and motivations.



The word cloud shows that employees are primarily concerned with the stigma, fear of negative consequences, and the potential impact of their mental health disclosure on hiring decisions. This underscores the need for creating safer, more supportive environments where mental health issues can be openly discussed without fear of discrimination or negative career repercussions.

The visualizations provided several insights into the mental health of tech employees:

- A significant number of respondents reported working for companies that do not offer mental health resources or support.
- Employees often expressed concern about discussing mental health issues at work due to fear of negative consequences.
- Anxiety and mood disorders were the most commonly reported mental health conditions, with significant gender differences in mental health outcomes.
- Mapping the respondents helped identify that most responses came from countries with developed tech industries, though some regions were underrepresented.

These visualizations were crucial for uncovering patterns in the dataset, and they served as the foundation for the subsequent predictive modeling and analysis of mental health outcomes in the tech industry.

4 Modeling

For this project, we explored a variety of machine learning models to predict whether an employee may have a mental health disorder based on their workplace environment and demographic factors. The models chosen were Random Forest, Logistic Regression, Gradient Boosting Machine (GBM), and Decision Tree, each offering distinct advantages for our analysis.

Before applying these models, categorical variables in the dataset were transformed using one-hot encoding. This method created binary variables for each category, allowing us to convert non-numeric data into a form that machine learning algorithms can process. By encoding the data in this way, we ensured that all models could handle the complexity of categorical features like workplace conditions, gender, and country, without assuming an ordinal relationship between the categories

Random Forest: Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees. Each tree is trained on a random subset of features and data, which helps prevent overfitting and improves the model's ability to generalize to new data. Given the complexity and interrelation of workplace factors and mental health disorders, Random Forest is well-suited for this task because it can handle a large number of input features and capture non-linear relationships between them. For our model, we set the number of trees (n_estimators) to 200, with a maximum depth (max_depth) of 7 to control overfitting, and a minimum sample size (min_samples_leaf) of 5 to ensure the trees aren't too specific to the training data. This model performed well with a test accuracy of 64% and provided valuable insights through feature importance, making it a strong candidate for our analysis. However, it faced challenges in predicting the "Maybe" class in our target variable, which indicates that uncertainty or partial mental health disorders are harder for the model to classify. Despite this limitation, Random Forest provided useful feature importance insights, making it the strongest model for our analysis.

Logistic Regression: Logistic Regression is a linear model often used for binary classification tasks, but in our case, it was applied to predict whether or not an employee has a mental health disorder across three possible outcomes. Logistic Regression works by estimating the probability that a given input belongs to a particular class. While simple, it is limited in its ability to capture complex interactions between features. In our study, we set the regularization parameter (C) to 1.5 and ran the model for up to 1000 iterations (max_iter). The test accuracy of Logistic Regression was lower compared to other models (53%), likely because mental health outcomes are influenced by more complex, non-linear relationships that the model could not fully capture. While it offered a clear interpretation of feature importance, its performance suggests that more advanced models are needed for this dataset.

Gradient Boosting Machine (GBM): Gradient Boosting is another ensemble learning technique, but unlike Random Forest, it builds models sequentially, where each new model attempts to correct the errors of the previous one. This allows GBM to handle complex datasets with higher accuracy. However, it can also be more prone to overfitting if not carefully tuned. For our model, we used 100 trees (n_estimators), a small learning rate (learning_rate = 0.001), and a randomized maximum depth to reduce overfitting. While the GBM achieved a test accuracy of 62%, its sequential nature and potential overfitting challenges made it less reliable for this dataset compared to Random Forest.

Decision Tree: Decision Trees are a simple, interpretable model that works by splitting the data based on the most important features at each level. It is easy to understand and visualize, which makes it a useful starting point for understanding feature interactions. However, Decision Trees can easily overfit if not constrained. To prevent this, we limited the maximum depth to 3. Although the Decision Tree achieved a test accuracy of 64%, comparable to Random Forest, it also struggled with correctly identifying the "Maybe" class. Its performance on more complex, non-linear data is limited compared to ensemble methods like Random Forest or Gradient Boosting.

In summary, Random Forest emerged as the best-performing model for this dataset, balancing interpretability, accuracy, and generalization capabilities. The ensemble approach allows for more robust predictions and a better handling of complex feature interactions, which are crucial in analyzing mental health outcomes based on workplace conditions.

All our models results are summarized in Table1.

The feature importance plot (Figure14) displays the top 10 most influential variables in the Random Forest model used to predict whether an employee has a mental health disorder. The most significant factor, by a wide margin, is whether the respondent has ever sought treatment for a mental health issue from a professional. This is confirmed also by the Random Forest Decision Tree plot (Figure16).

Other key features include the presence of family history of mental illness, both positive ("Yes") and negative ("No"), which further reinforces the influence of personal and family mental health backgrounds on predicting current mental health status.

Experiences related to unsupportive or badly handled responses to mental health issues in the workplace also play a crucial role. This suggests that negative workplace experiences surrounding mental health discussions may exacerbate or reflect existing conditions.

Further down, we see factors like whether team members or co-workers would view the respondent negatively if they knew about their mental health disorder, and whether the employer provides mental health benefits. These indicate that perceptions of workplace stigma and the availability of mental health support are also important in determining an individual's mental health status.

5 Conclusions

Based on our analysis, we have drawn several conclusions regarding how to better support employees who may be experiencing mental health challenges. While our recommendations are primarily directed at the HR departments of tech companies, they are equally applicable to colleagues, supervisors, and managers. Implementing these strategies can help foster a more supportive and inclusive workplace for everyone.
Possible recommendations are here defined;

- **Increase Awareness and Reduce Stigma Around Mental Health**; Implement comprehensive mental health awareness programs that foster open discussions about mental health issues in the workplace. It could be useful the introduction a *mental health ambassador program*, where employees can voluntarily become advocates for mental well-being in the workplace.

- **Mental Health Resources and Access to Professional Help**; Ensure that employees have access to adequate mental health benefits, including therapy, counseling, and professional treatment. Partner with mental health professionals and services to provide confidential and free or subsidized counseling for employees.

- **Create a Supportive Work Environment**; Foster a work environment where employees feel safe to disclose mental health issues without fear of discrimination. For example, Microsoft actively encourages conversations about mental health by integrating mental well-being into its Diversity & Inclusion strategy.
  Another possible idea is creating a more inclusive and supportive workplace environment by organizing family-inclusive events where employees can invite their families. We define it as the "*Humanizing the Workplace*". Introducing families into the work environment allows employees to see each other in a more personal and empathetic context. This reduces the "formal" barriers that often prevent open discussions around sensitive topics like mental health.

- **Promote Work-Life Balance to Mitigate Stress**; Address workplace stress by promoting flexible work arrangements and encouraging a healthy work-life balance.

- **Use Data to Continuously Improve Mental Health Support**; the usage of analysed data, such as our model, can be useful to identify trends. HR departments can implement an ongoing feedback loop, where they gather data from employees about their mental health experiences and how workplace conditions impact their well-being.

By implementing these recommendations, HR departments can help build a healthier, more supportive, and inclusive workplace that prioritizes mental health, leading to improved employee well-being and organizational success. These solutions not only address mental health concerns directly but also set a foundation for long-term cultural change in how mental health is viewed and managed in the tech industry.

Table 1: Model Performance Comparison

Model	**Training Accuracy**	**Test Accuracy**
Random Forest	70%	64%
Logistic Regression	71%	53%
Gradient Boosting	64%	62%
Decision Tree	65%	64%

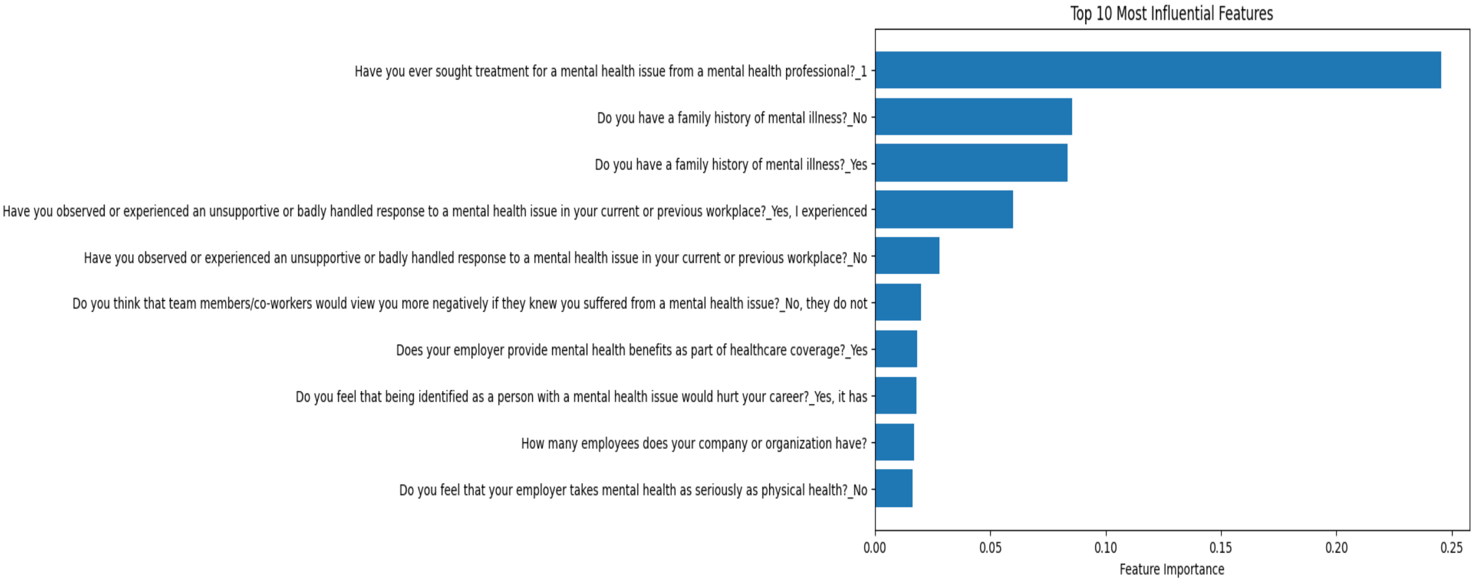


Figure 14: Most Important Features in Random Forest

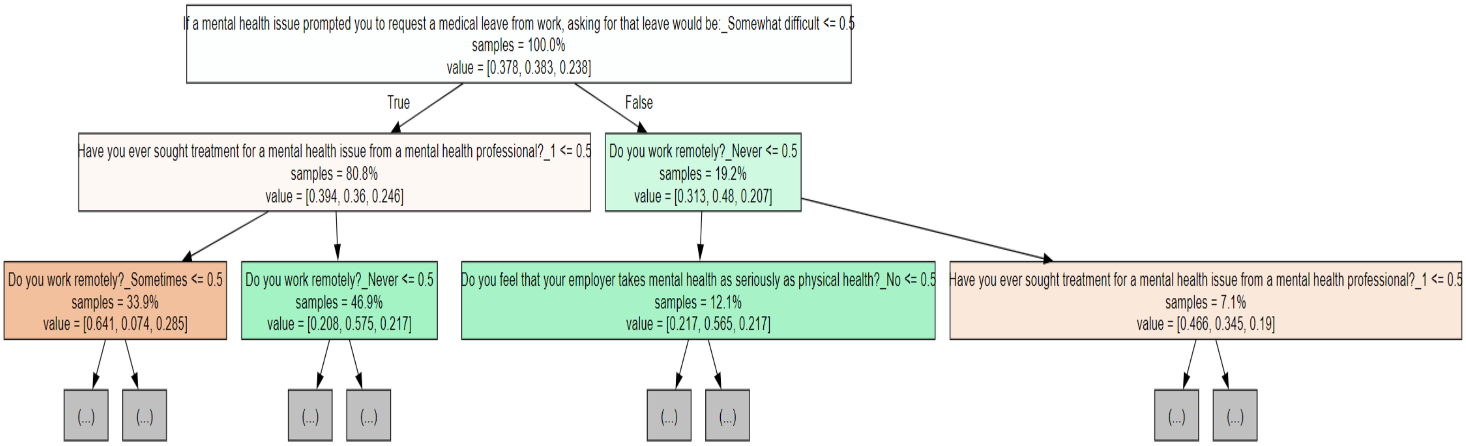

If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:_Somewhat difficult <= 0.5
samples = 100.0%
value = [0.378, 0.383, 0.238]
True
False
Have you ever sought treatment for a mental health issue from a mental health professional?_1 <= 0.5
samples = 80.8%
value = [0.394, 0.36, 0.246]
Do you work remotely?_Never <= 0.5
samples = 19.2%
value = [0.313, 0.48, 0.207]
Do you work remotely?_Sometimes <= 0.5
samples = 33.9%
value = [0.641, 0.074, 0.285]
Do you work remotely?_Never <= 0.5
samples = 46.9%
value = [0.208, 0.575, 0.217]
Do you feel that your employer takes mental health as seriously as physical health?_No <= 0.5
samples = 12.1%
value = [0.217, 0.565, 0.217]
Have you ever sought treatment for a mental health issue from a mental health professional?_1 <= 0.5
samples = 7.1%
value = [0.466, 0.345, 0.19]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Figure 15: Random Forest First Decision Tree output

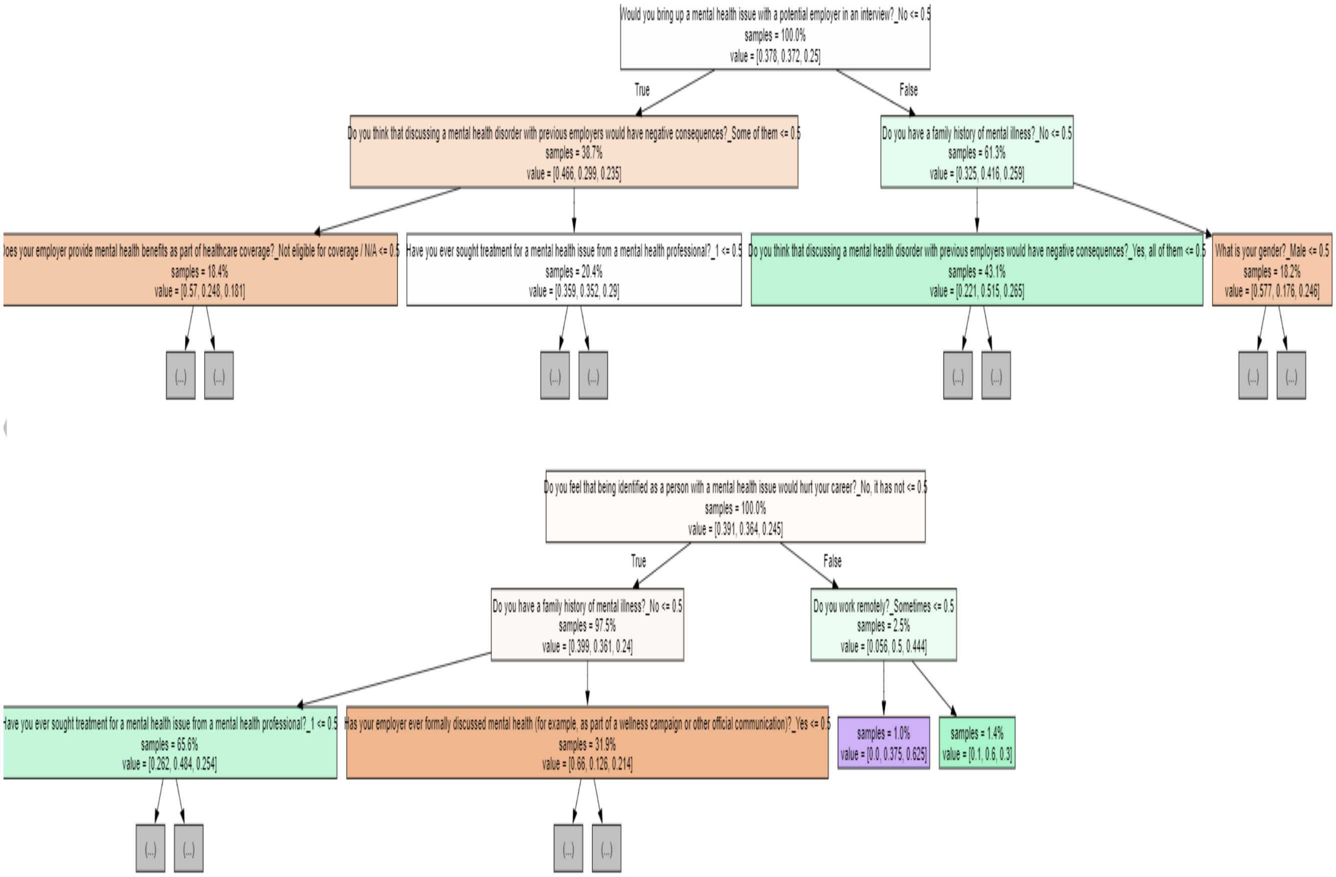

Would you bring up a mental health issue with a potential employer in an interview?_No <= 0.5
samples = 100.0%
value = [0.378, 0.372, 0.25]
True
False
Do you think that discussing a mental health disorder with previous employers would have negative consequences?_Some of them <= 0.5
samples = 38.7%
value = [0.466, 0.299, 0.235]
Do you have a family history of mental illness?_No <= 0.5
samples = 61.3%
value = [0.325, 0.416, 0.259]
oes your employer provide mental health benefits as part of healthcare coverage?_Not eligible for coverage / N/A <= 0.5
samples = 18.4%
value = [0.57, 0.248, 0.181]
Have you ever sought treatment for a mental health issue from a mental health professional?_1 <= 0.5
samples = 20.4%
value = [0.359, 0.352, 0.29]
Do you think that discussing a mental health disorder with previous employers would have negative consequences?_Yes, all of them <= 0.5
samples = 43.1%
value = [0.221, 0.515, 0.265]
What is your gender?_Male <= 0.5
samples = 18.2%
value = [0.577, 0.176, 0.246]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
Do you feel that being identified as a person with a mental health issue would hurt your career?_No, it has not <= 0.5
samples = 100.0%
value = [0.391, 0.364, 0.245]
True
False
Do you have a family history of mental illness?_No <= 0.5
samples = 97.5%
value = [0.399, 0.361, 0.24]
Do you work remotely?_Sometimes <= 0.5
samples = 2.5%
value = [0.056, 0.5, 0.444]
ave you ever sought treatment for a mental health issue from a mental health professional?_1 <= 0.5
samples = 65.6%
value = [0.262, 0.484, 0.254]
Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?_Yes <= 0.5
samples = 31.9%
value = [0.66, 0.126, 0.214]
samples = 1.0%
value = [0.0, 0.375, 0.625]
samples = 1.4%
value = [0.1, 0.6, 0.3]
(...)
(...)
(...)
(...)

Figure 16: Random Forest Second and Third Decision Tree outputs

Acknowledgments

We would like to thanks Chat-GPT for enhancing the clarity and quality of the written content in this report.

Credit Author Statement

Bozhok Yaroslava; Topic Decision, Methodology, Dataset Investigation, Presentation structure

Makarov Valeri; Topic Decision, Methodology, Data Visualization, Modeling, Insights

Filippo Navarra; Topic Decision, Methodology, Data Visualization, EDA, Modeling, Insights, Report, Presentation, Supervision.