Analyze Mental health disorder in Tech Companies

DSC550-T301 | CHITRAMOY MUKHERJEE | DATE : 02/026/2024

1. Introduction:

In recent years, the tech industry has experienced rapid growth and innovation, bringing about numerous opportunities and challenges. While technological advancements have transformed the way we work, they have also introduced new stressors that can impact the mental health of individuals working in this sector. Recognizing the importance of mental health in the workplace, this project aims to analyze mental health disorders within tech companies using Python. Mental health affects your emotional, psychological and social well-being. Mental health is a key factor in determining the employee's productivity in any industry and the total performance of the company. If someone is not mentally fit, he can't produce the expected output what he is capable of, and it also impacts his co-worker's performance and impacts the work environment.

By the end of this project, we aim to contribute valuable insights that can inform both employers and employees about mental health in the tech industry. This analysis can serve as a foundation for fostering a healthier and more supportive work environment both from employee and employer perspective. Employers can also offer robust benefit packages to support employees who go through mental health issues. That includes Employee Assistance Programs, Wellness programs that focus on mental and physical health, Health and Disability Insurance or flexible working schedules or time off policies. Below are the key factors why addressing mental health in the tech industry is crucial:

Employee Well-being: Prioritizing mental health fosters a supportive and healthy work environment. Employees spend a substantial amount of time at work, and their well-being directly impacts job satisfaction and overall life quality.
Productivity and Performance: Mental health issues can negatively impact job performance, creativity, and productivity. By addressing mental health concerns, employees are more likely to be engaged, focused, and perform at their best, contributing to the overall success of the company.
Retention and Recruitment: A company that actively supports mental health is likely to attract and retain top talent. Employees value employers who prioritize their well-being and create a culture that promotes work-life balance and mental health support.
Awareness and Education: Providing resources to raise awareness about mental health, reduce stigma, and educate employees on recognizing signs of distress.
Supportive Workplace Culture: Fostering a workplace culture that prioritizes work-life balance, stress management, and open communication about mental health challenges.

Integration into Benefits Package: Including mental health support as an integral part of our employee benefits package, demonstrating our commitment to employee well-being.
Reduced Absenteeism: Mental health disorders can lead to absenteeism due to sick days or other related issues. By addressing mental health concerns, companies may experience reduced absenteeism, which contributes to a more consistent and reliable workforce.
Innovation and Creativity: A positive mental health culture fosters creativity and innovation. Employees who feel supported and valued are more likely to contribute new ideas and solutions, driving innovation within the organization.
Diversity and Inclusion: Promoting mental health awareness and support contributes to creating an inclusive workplace culture. Recognizing and addressing mental health challenges can help break down stigmas and foster an environment where diverse perspectives are valued.
Legal and Ethical Responsibility: Companies have a legal and ethical responsibility to provide a safe and healthy working environment. Neglecting mental health concerns may lead to legal implications, and it is in the best interest of companies to adhere to workplace health and safety standards.
Cost Savings: Addressing mental health issues early can result in cost savings for companies. Investing in employee well-being programs and mental health resources can lead to decreased healthcare costs, reduced turnover, and increased productivity.
Positive Company Reputation: Companies that prioritize mental health are more likely to be viewed favorably by employees, customers, and the public. This positive reputation can enhance the company's brand image and attract customers who value socially responsible business practices.

This topic is relevant to data science as we can analyze and identify the factors/variables that impact mental health and justify the relations between variables which are closely related to determine the mental health of employees. We can create a model and feed data into it to identify the employee's mental health in the company and provide directions to them to overcome the situation.

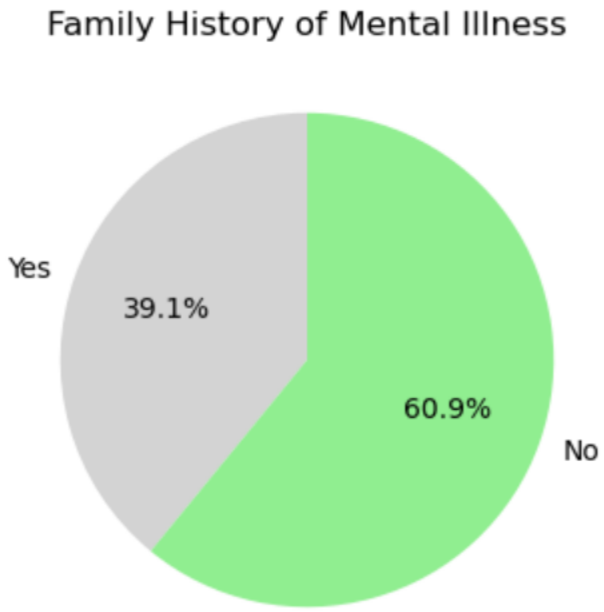
This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. For this analysis this dataset is sourced from kaggle and it contains the Data between August,2014 to Feb,2016.

2. Organized and detailed summary of Milestones 1-3 :

- **EDA :**

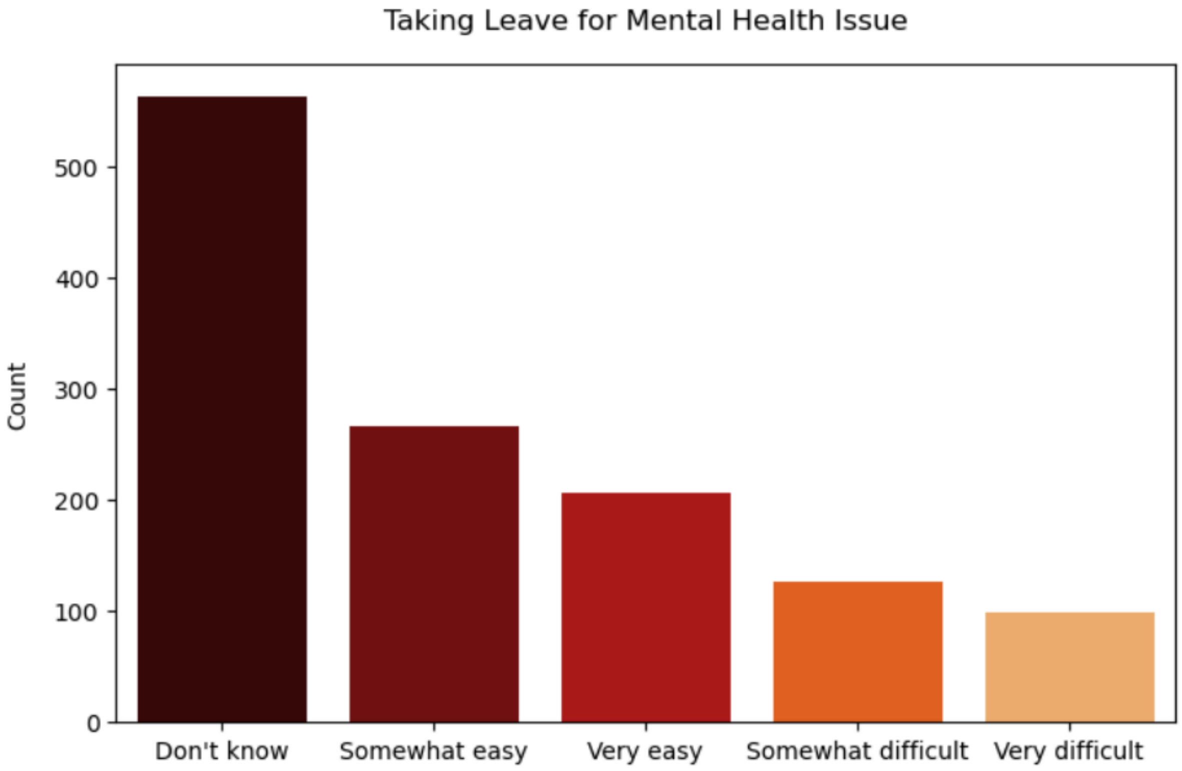
EDA (Exploratory Data Analysis) plays a crucial role in the machine learning pipeline by providing a comprehensive understanding of the data, guiding feature selection and engineering, and helping in the overall improvement of model performance and reliability. Below are the plotting's performed on data during EDA.

Pie Diagram on Family History of Mental health:

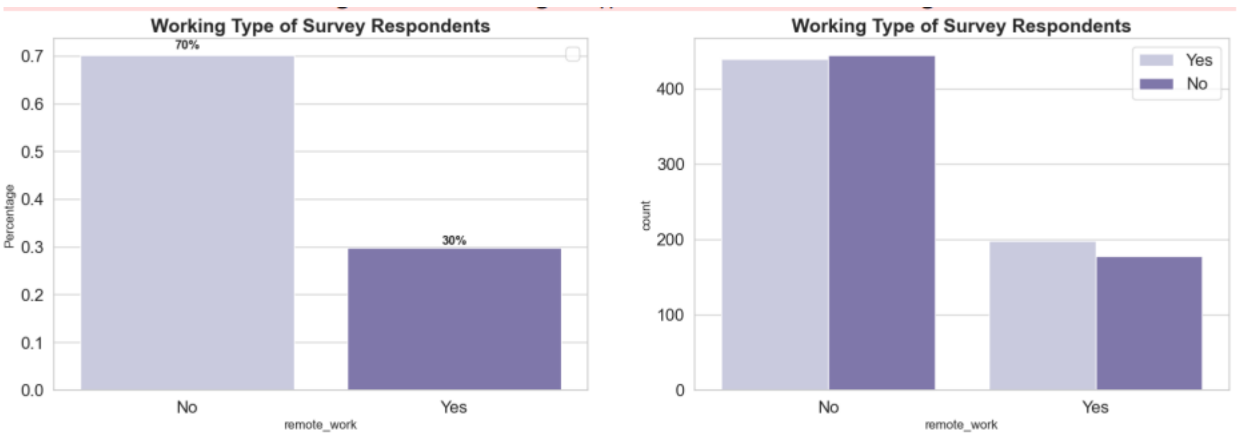


From this, we can see that almost 40% of respondents have a family history of mental illness. According to a 2017 study by the Arctic University of Norway, it was discovered that children with parents who had a severe mental illness had up to a 50% chance of developing a mental illness, and a 32% chance of developing a severe mental illness (bipolar disorder, major depressive disorder, schizophrenia, etc). We will look further into this when performing bivariate analysis.

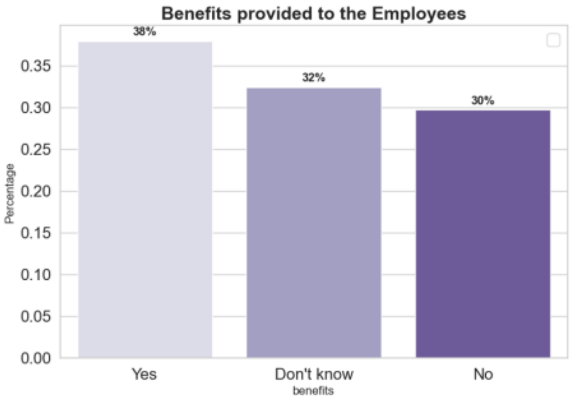
- **Bar diagram on taking leave for Mental health issue :**

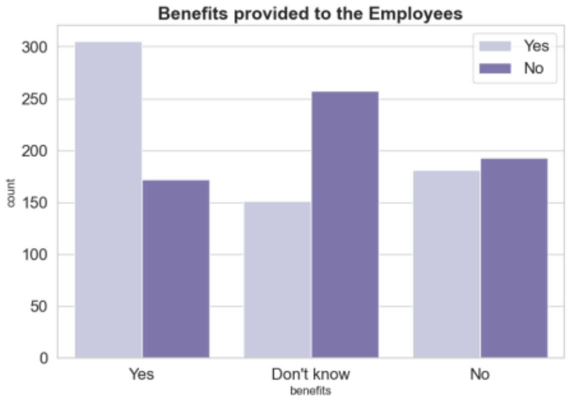


From the above plot, we can see that most respondents do not know whether they are even allowed to take leave for a mental health issue, and there are also quite a number who find it hard to do so, which may be due to the social stigma surrounding mental issues.



Around 70% of respondents don't work remotely, which means the biggest factor of mental health disorder came up triggered on the workplace. On the other side, it has slightly different between an employee that want to get treatment and don't want to get a treatment. The number of people who seek treatment in both the categories is more or less similar and it does not affect our target variable.



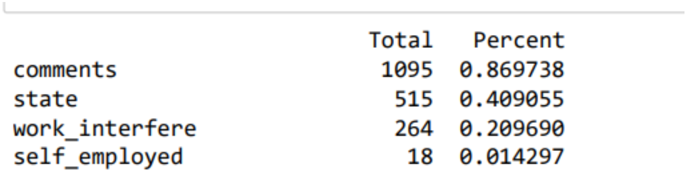


We see that around 38% of the respondents said that their employer provided them mental health benefits, whereas a significant number (32%) of them didn't even know whether they were provided this benefit. Coming to the second graph, we see that for the people who YES said to mental health benefits, around 63% of them said that they were seeking medical help. Surprisingly, the people who said NO for the mental health benefits provided by the company, close to 45% of them who want to seek mental health treatment.

- **Data preparation:**

First step during data preparation was visualizing the data and identifying the non-null values. Below are the steps performed for data preparation.

1. Identified below 4 columns which have Null values.



```
                 Total   Percent
comments          1095  0.869738
state              515  0.409055
work_interfere     264  0.209690
self_employed       18  0.014297
```

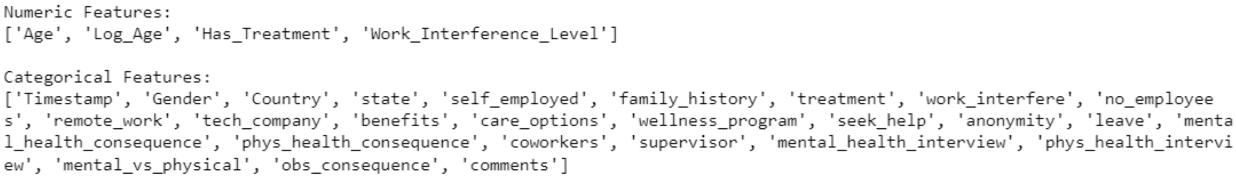
2. Timestamp, Comments, State are the columns which are being dropped as those are not columns required for analysis.
3. Next step was to assign default values for each data type and Clean the NaN's with some default values.
4. Standardized the Gender column value into 'female' 'male' 'trans' .
5. Completed the missing age with the mean value.
6. Only 0.014% of self-employed so changed NaN to NOT self_employed.
7. 0.20% of self work_interfere so let's change NaN to "Don't know".

Once the basic cleanup of data was performed, next step was to identify the features and target variables and derive the new features for modeling.

Derive below 3 new features,

1. Age Group : A new categorical feature is created to represent different age groups based on the 'Age' column.
2. Has Treatment: A binary feature is created indicating whether the respondent has received treatment or not.
3. Work Interference Level: A numerical feature is created to represent the level of work interference, mapping categorical values to numerical values.

Identify the Numerical and Categorical features from the dataset for modeling.



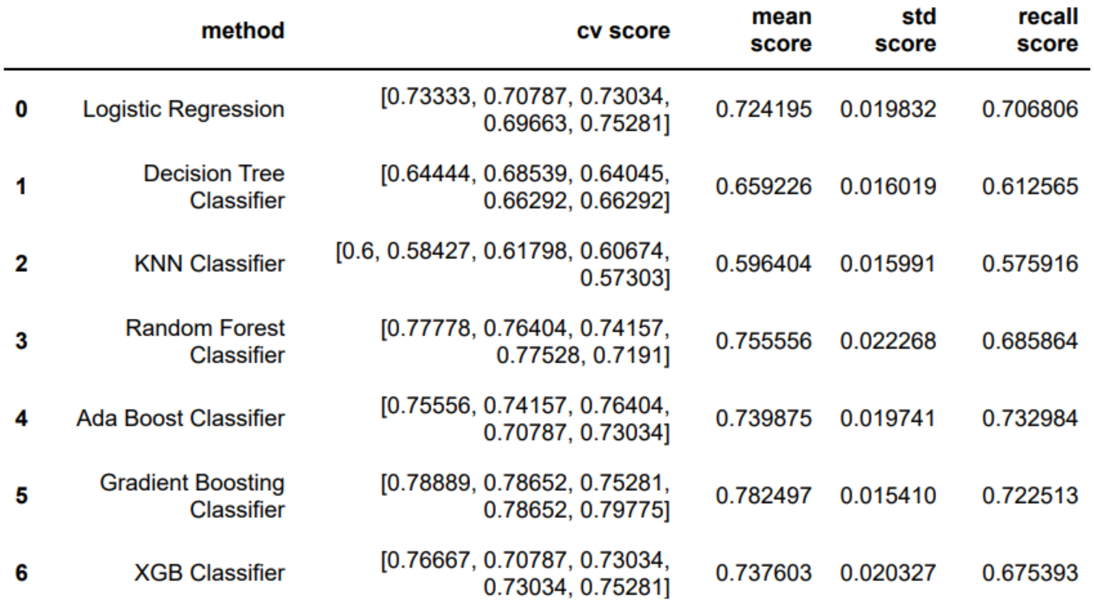
```
Numeric Features:
['Age', 'Log_Age', 'Has_Treatment', 'Work_Interference_Level']

Categorical Features:
['Timestamp', 'Gender', 'Country', 'state', 'self_employed', 'family_history', 'treatment', 'work_interfere', 'no_employee
s', 'remote_work', 'tech_company', 'benefits', 'care_options', 'wellness_program', 'seek_help', 'anonymity', 'leave', 'menta
l_health_consequence', 'phys_health_consequence', 'coworkers', 'supervisor', 'mental_health_interview', 'phys_health_intervi
ew', 'mental_vs_physical', 'obs_consequence', 'comments']
```

- **Model building and evaluation :**

As a part of modeling exercise, 3 basic models (Logistic Regression, KNN and Decision Tree Classifier) and 4 ensembles models (Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, XGB Classifier). Below are the cv scroe, mean score, std score and recall score calculated for different modeling method.

From the below cross validation process, there are 2 models that pop up with high precision scores. The first is Logistic Regression for the basic model and the second is Ada Boost Classifier for the ensemble model. Considering Logistic Regression because Ada Boost Classifier is really heavy to process.



	method	cv score	mean score	std score	recall score
0	Logistic Regression	[0.73333, 0.70787, 0.73034, 0.69663, 0.75281]	0.724195	0.019832	0.706806
1	Decision Tree Classifier	[0.64444, 0.68539, 0.64045, 0.66292, 0.66292]	0.659226	0.016019	0.612565
2	KNN Classifier	[0.6, 0.58427, 0.61798, 0.60674, 0.57303]	0.596404	0.015991	0.575916
3	Random Forest Classifier	[0.77778, 0.76404, 0.74157, 0.77528, 0.7191]	0.755556	0.022268	0.685864
4	Ada Boost Classifier	[0.75556, 0.74157, 0.76404, 0.70787, 0.73034]	0.739875	0.019741	0.732984
5	Gradient Boosting Classifier	[0.78889, 0.78652, 0.75281, 0.78652, 0.79775]	0.782497	0.015410	0.722513
6	XGB Classifier	[0.76667, 0.70787, 0.73034, 0.73034, 0.75281]	0.737603	0.020327	0.675393

Calculated the best score for logistic regression before and after tunning and below is the output.

Let's discuss the implications of the output metrics:

- Method (Algorithm):
  These are the machine learning algorithms being evaluated: Logistic Regression, Decision Tree Classifier, KNN Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, and XGB Classifier.
- CV Score (Cross-Validation Score):
  This represents the recall score for each algorithm obtained through cross-validation. Recall is a measure of a model's ability to correctly identify positive instances. A higher recall indicates better performance in identifying positive cases.
- Mean Score:
  This is the mean recall score across the folds of cross-validation. It provides a summary measure of the overall performance of each algorithm. A higher mean score suggests better average recall performance.
- Std Score (Standard Deviation of Cross-Validation Score):
  The standard deviation of the recall scores provides a measure of the variability or consistency of the model's performance across different folds. Lower standard deviation implies more consistent performance.
- Recall Score:

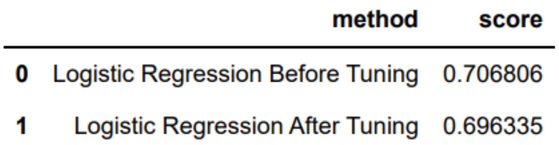
This is the recall score on the test set for each algorithm. It gives an indication of how well the model generalizes to new, unseen data.

- **Definitions of different algorithms :**

  - Logistic Regression:
    It is used for binary classification and provides a baseline performance.
  - Decision Tree Classifier:
    Can capture complex relationships but may be prone to overfitting.
  - KNN Classifier:
    Sensitive to outliers and requires careful preprocessing.
  - Random Forest Classifier:
    Reduces overfitting by combining multiple decision trees.
  - Ada Boost Classifier:
    Emphasizes misclassified samples, potentially improving performance.
  - Gradient Boosting Classifier:
    Builds models sequentially, correcting errors of the previous ones.
  - XGB Classifier:
    An optimized version of gradient boosting, often delivering high performance.

  - Mean Score and Std Score:
    Look for algorithms with higher mean scores and lower standard deviation, indicating consistent and good overall performance.
  - Recall Score:
    The recall on the test set indicates the model's ability to correctly identify positive instances in real-world scenarios.

Consideration of these metrics will help in choosing the algorithm that best suits the specific requirements of the problem, balancing performance and generalization. Additionally, further tuning and optimization may be applied to enhance the model's performance.



	method	score
0	Logistic Regression Before Tuning	0.706806
1	Logistic Regression After Tuning	0.696335

After Feature Selection Process, the score is 0.6963350785340314.

Conclusion :

By considering all these values, you get a comprehensive understanding of how well each algorithm is likely to perform on unseen data and how consistent its performance is across different training subsets. This information helps in justifying the accuracy of the model and selecting the most suitable algorithm for the given task.

- **CV Score:** Provides a detailed breakdown of performance across different subsets of data.
- **Mean Score:** Offers a single, summary metric for overall performance.
- **Std Score:** Indicates the consistency or variability in performance.
- **Recall Score:** Reflects the model's ability to identify positive instances on new data.

We have calculated various scores on machine learning algorithm's and then based on the score we have evaluated Logistic Regression and the score for logistic regression after and before tunning is around 70%. Since correlation between variables is very important in a regression project, perhaps the required correlation was not present in this dataset. I think the use of clustering and PCA or even SOM can give a good result in this data set.

Is this Model is ready to be deployed:

Logistic Regression model has high CV scores, a respectable mean score, low standard deviation, and a good recall score on the test set, it suggests that the model is performing well in terms of recall and generalization.

However, the decision to deploy a model should also take into account other factors, such as the nature of the problem, interpretability of the model, business requirements, and ethical considerations. Additionally, thorough testing on real-world data and potential monitoring mechanisms post-deployment are essential steps.

Recommendations:

To provide recommendations on the model and analysis, let's analyze the provided information and consider best practices:

- Consider Model Performance:

  Evaluate each model based on the CV scores, mean score, standard deviation, and recall score.

Look for a balance between high recall on the test set and consistency across cross-validation folds.

Compare the performance of different models to identify the most suitable one.

- Tune Hyperparameters:

  Consider hyperparameter tuning for the selected model to potentially improve performance.

  Utilize techniques like GridSearchCV or RandomizedSearchCV for hyperparameter optimization.

- Feature Importance:

  Investigate the importance of features in the chosen model.

  Understanding feature importance can provide insights into which features contribute significantly to the model's predictions.

- Handle Class Imbalance:

  If your dataset is imbalanced, assess whether the model handles class imbalance effectively.

  Techniques such as oversampling, undersampling, or using specialized algorithms can be considered.

- Deploy the Best Model:

  Once a model is selected, deploy it in a production environment.

  Consider the implications of the model in a real-world setting, such as interpretability, scalability, and computational efficiency.

- Monitoring and Maintenance:

  Implement monitoring mechanisms to track the model's performance over time.

  Regularly update the model if new data becomes available or if the distribution of the data changes.

- Ethical Considerations:

  Examine the ethical implications of the model, especially if it is used in decision-making processes.

  Address any potential biases and ensure fairness.

- Continuous Improvement:

  Continue to iterate and improve the model based on new insights and data.

  Embrace a continuous improvement mindset in the deployment and maintenance phases.

Suitability of the recommendations depends on the specific characteristics of your dataset, problem domain, and business context. Regular validation and collaboration with domain experts and stakeholders are essential for a successful deployment.

while the provided scores are important indicators, a comprehensive assessment considering various aspects is necessary before deciding whether the Logistic Regression model is ready for deployment.

Challenges:

- Data Privacy and Ethics:
  Mental health data is sensitive, and maintaining privacy is crucial. Ensuring ethical data collection, handling, and storage is a challenge in this context.
- Data Availability:
  Access to comprehensive mental health datasets in the tech industry may be limited. Gathering representative and diverse data could be challenging.
- Stigma and Self-Reporting Bias:
  Mental health issues often come with stigma. Employees may be hesitant to disclose their mental health status, leading to potential biases in self-reported data.
- Multifactorial Nature of Mental Health:
  Mental health is influenced by various factors, including work environment, personal life, and genetics. Capturing and analyzing these multifaceted aspects is complex.
- Interdisciplinary Collaboration:
  Addressing mental health issues requires collaboration between data scientists, mental health professionals, HR experts, and other stakeholders. Building effective interdisciplinary teams can be challenging.
- Imbalanced Datasets:
  Imbalances in the prevalence of mental health conditions may lead to biased models. Techniques to handle imbalanced datasets need consideration.

Opportunities:

- Predictive Modeling for Early Intervention:
  Develop models to predict the risk of mental health issues early, allowing for proactive intervention and support.
- Natural Language Processing (NLP):
  Analyze text data, such as employee feedback, emails, or chat messages, using NLP to detect sentiment and identify potential mental health concerns.
- Employee Well-Being Platforms:
  Integrate data-driven insights into well-being platforms to provide personalized support and resources for employees.

- Incorporate Wearable and Sensor Data:

  Utilize wearable devices and sensor data to monitor physiological i indicators that may be correlated with mental health states.
- Examine Work-Life Balance:

  Explore the impact of work-related factors, such as workload, deadlines, and work culture, on mental health. Identify patterns that contribute to stress or well-being.
- Examine Remote Work Challenges:

  Investigate the mental health challenges associated with remote work, considering isolation, digital communication stress, and work-life integration.
- Cultural and Gender Considerations:

  Examine how cultural and gender factors influence mental health experiences in the tech industry. Tailor interventions accordingly.
- Explainability and Interpretability:

  Develop models that provide interpretable results, allowing stakeholders to understand the factors influencing mental health predictions.