Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	738214
Project Title	Predicting Mental Health Illness Of Working Professionals Using Machine Learning.
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to leverage machine learning to develop a model that can predict mental health risk among tech industry workers. By analyzing data from a survey of tech professionals, we hope to gain valuable insights into factors associated with mental health challenges in this specific population. **Project Objectives:** 1. **Identify Key Risk Factors:** Explore the survey data to identify key features (variables) that are most strongly associated with mental health outcomes. 2. **Build Predictive Model:** Develop a machine learning model that can predict the likelihood of experiencing mental health issues based on these identified risk factors. 3. **Improve Mental Health Support:** Ultimately, by understanding the factors that contribute to mental health challenges in the tech industry, we can inform interventions and support systems to promote better mental well-being among tech workers.

	Potential Benefits: • Early identification of mental health risks can lead to early intervention and improved outcomes. • The model can be used to develop targeted mental health resources and support programs for tech companies. • The findings can inform broader discussions about mental health awareness and create a more supportive work environment for tech professionals. **About Data:** To crack the case of mental health challenges in tech workers, we'll be like detectives! Our evidence comes from a survey of tech professionals, where questions about demographics, work experiences, and mental well-being act as our clues. By analyzing this data (age, stress levels, work-life balance), we aim to build a model that predicts who might be at risk, ultimately helping us develop better support systems for this industry. **Machine Learning Approach (to be determined):** The specific machine learning algorithms used will depend on the nature of the data and the desired outcomes. However, potential approaches include: • **Logistic Regression:** For classifying individuals into high or low risk categories for mental health problems. • **Random Forests:** To capture complex relationships between multiple features and mental health outcomes.
Data Collection Plan	The "Mental Health in Tech Survey" dataset provides information about attitudes and experiences related to mental health in the tech industry. This data comes from a 2014 survey that explores employee demographics (age, gender, location) and work characteristics (company size, remote work, benefits). It delves deeper into mental health aspects by asking if participants have a family history of mental illness or have sought treatment themselves. The survey also investigates how mental health is perceived in the workplace, including whether employees feel comfortable discussing mental health issues with supervisors, colleagues, or even during job interviews. Additionally, it explores company practices such as offering mental health resources and

	programs, and employee perceptions of the consequences of disclosing mental health concerns. This rich dataset offers valuable insights into the intersection of mental health and the tech industry workforce.
	
	Primary Data Source:
	
	• This project will primarily utilize the existing "Mental Health in Tech Survey" dataset. Here's why:
	  ○ It directly addresses your project's goals of understanding mental health in the tech industry.
	  ○ The data offers a rich set of features relevant to your analysis (demographics, work environment, mental health perceptions).
	  ○ Utilizing existing data saves time and resources compared to conducting a new survey.
	
	Data Acquisition:
	
	• Download the 2014 "Mental Health in Tech Survey" dataset from the provided source.
	• If using the 2016 survey, follow the instructions provided on the website (mentioned as "found here" but not provided).
	
	Data Quality Considerations:
	
	• Since you're using existing data, assess its quality by checking for:
	
	  ○ Missing values and data cleaning procedures needed.
	  ○ Documentation about data collection methods and potential biases.
	  ○ Sample size and representativeness of the tech industry population.
	
	Ethical Considerations:
	
	• Even though the data is anonymized, emphasize respecting participant privacy in your project documentation.
	
	Next Steps:
	
	• Once you have the data, explore its features and begin your analysis to identify key factors related to mental health in the tech industry.

Raw Data Sources Identified	**Mental Health in Tech Survey (2014):** This is the primary data source for the project. It's a survey conducted in 2014 that explores mental health attitudes and experiences within the tech industry workforce. The data includes demographics (age, gender, location), work characteristics (company size, remote work, benefits), mental health history and treatment, workplace perceptions of mental health, and company practices related to mental health resources and support. **Optional: Mental Health in Tech Survey (2016):** This can be a secondary data source if available. It would provide data from a more recent survey on the same topic, allowing for comparisons over time or potentially offering a larger or more representative sample. However, access to this data source might require further investigation based on the provided information.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Mental Health in Tech Survey	This is the primary data source for the project. It's a survey conducted in 2014 that explores mental health attitudes and experiences within the tech industry workforce. The data includes demographics (age, gender, location), work characteristics (company	https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey	CSV	303.68 kB	Public

	size, remote work, benefits), mental health history and treatment, workplace perceptions of mental health, and company practices related to mental health resources and support.				
Mental Health	The importance of mental health and how common mental health conditions are. It then dives into how surveys are used to assess mental health. Surveys offer valuable insights because they ask standardized questions, reaching a wider population than those seeking professional help. However, limitations exist as people might be hesitant to disclose symptoms, and accurately recalling past experiences can be challenging.	https://www.kaggle.com/datasets/imtkaggleteam/mental-health	CSV	451.49 KB	Public