Project title: Mental Health in Tech

Team members: Svend Sakkool, Tambet Osman

Task 2: Business understanding

1. Background and business context

The technology sector shows growing attention to employee mental health, but many questions remain about who seeks treatment and which workplace or personal factors influence that decision. Using the publicly available Kaggle survey “Mental Health in Tech” as our primary data source, this project aims to combine exploratory analysis, pattern discovery, and predictive modelling to understand and forecast treatment-seeking behaviour among tech employees. The exercise is primarily educational — intended to develop practical data-analysis skills — but the findings will also be of interest to tech companies and stakeholders concerned with workforce wellbeing.

2. Business goals

- Produce an interpretable, data-driven characterisation of tech employees who sought professional mental health treatment versus those who did not.
- Identify common demographic, workplace, and attitudinal factors associated with treatment-seeking.
- Build a predictive model capable of flagging individuals with a high probability of seeking treatment, to demonstrate how analytics could inform targeted wellbeing interventions.
- Deliver clear visualisations and actionable recommendations that could be communicated to industry stakeholders.

3. Business success criteria

Success will be judged by a combination of analytical quality and communicability:

- A classification model with reasonable discrimination (target: AUC significantly above random; exact numeric goal to be refined during modelling).
- Robust, reproducible exploratory analyses (visualisations + correlation summaries) that clearly surface key predictors.
- Interpretable outputs (feature importance, rules or profiles from association/clustering analyses) that support at least three concrete, data-backed recommendations for employers or researchers.

- Deliverables: a clean Jupyter notebook, summary report, and a short slide deck with top findings and visuals.

4. Assessing the situation

Inventory of resources

- Data: Kaggle dataset “Mental Health in Tech” (survey responses).
- Tools: Python (pandas, scikit-learn, mlxtend or similar for association rules, clustering libraries, visualization libraries), Jupyter Notebook.
- Skills: Team experience with exploratory analysis, classification, clustering and pattern mining.

Requirements, assumptions and constraints

- Requirement: all analysis must be reproducible and documented in notebooks.
- Assumption: survey responses are self-reported and representative enough for pattern discovery (not necessarily for causal claims).
- Constraint: single dataset only (no internal company data), so generalisability is limited.

Risks and contingencies

- Self-selection and reporting bias in survey responses may limit external validity — results will be framed as exploratory and descriptive.
- Class imbalance (fewer respondents reporting treatment) could impair predictive performance — address with resampling, calibrated probabilities, and careful metric selection.
- Missing values and inconsistent coding will require cleaning; contingency: document imputation strategies and sensitivity checks.

Costs and benefits

- Low monetary cost (public dataset, open-source tools). Main cost: team time.
- Benefits: hands-on learning, clear visual and model outputs, and insights that may inform employer wellbeing discussions.

5. Data-mining goals

- Exploratory data analysis and visualization to summarize distributions and relationships.
- Association rule mining to discover frequent co-occurring attributes among those who sought treatment.

- Clustering to identify respondent segments with distinct risk/protection profiles.
- Supervised classification to predict treatment-seeking and quantify driver importance (e.g., logistic regression, tree-based models, with post-hoc interpretation like feature importance or partial dependence).

6. Data-mining success criteria

- EDA: clear visual summaries covering key variables and missingness.
- Association rules: stable, interpretable rules with reasonable support/confidence.
- Clusters: cohesive and meaningful clusters validated by internal metrics and described with characteristic features.
- Classification: models with reliable discrimination and calibration; interpretability prioritized (feature weights, importance, or model-agnostic explanations). Final acceptance will consider both predictive performance and interpretability.

Task 3. Data understanding

1. **Gathering data**
    - Objective: Analyze mental health in the tech workplace.
    - Data Required: Employee demographics, mental health history, workplace support, treatment behavior
    - Availability: Kaggle dataset Mental Health in Tech Survey (focus on 2016).
    - Selection Criteria: Include respondents with key mental health and workplace data; exclude blank or irrelevant entries.
2. **Describing data**
    - Format: CSV, ~1,400 rows, 27 attributes.
    - Types:
    - Categorical: gender, country, benefits, treatment, remote work, etc.
    - Numerical: age, number of employees.
    - Examples:
    - Age: numeric values (validated range 15–120).
    - Gender: highly inconsistent free-text values.
    - treatment: Yes/No responses.

3. **Exploring data**

- Distributions: Mostly males, age 20–35, mainly US/Canada/UK respondents. There were also some Non-Binary individuals in the dataset, though in much smaller numbers. The majority of respondents reported working in tech companies or remotely.
- Relationships: Work interference often appears linked to seeking treatment; mental health benefits may reduce stigma. However, these patterns need further analysis before drawing conclusions, as additional statistical testing is required to confirm any correlations.

4. **Verifying data quality**

Issues identified:

- Missing values in several key fields.
- Extremely inconsistent Gender entries (e.g., “male-ish”, “cis man”, “queer/she/they”).
- Age values outside realistic limits.
- Mixed formats for “don’t know” responses ("Unsure/Missing", "Not sure", "Don't know").
- Yes/No values stored as strings.
- Ordinal fields stored as text (e.g., work_interfere and no_employees).
- Non-US respondents had meaningless state codes.
- Some survey answers included vague responses such as “Maybe,” “Some of them,” or “Very difficult,” which required consolidation.

Actions Taken :

- Converted timestamps to proper datetime format.
- Standardized gender entries by mapping dozens of free-text terms into Male, Female, and Non-Binary; invalid entries set to NaN.
- Cleaned age values and removed impossible entries (<15 or >120).
- Set all non-US state values to "N/A" and filled missing states with "Missing".
- Standardized uncertain responses by replacing "Don’t know" and "Not sure" with "Unsure/Missing".
- Encoded work_interfere using an ordinal scale:

Never = 0

Rarely = 1

Sometimes = 2

Often = 3

- Encoded company size categories into ordered numerical values.
- Converted multiple Yes/No variables into binary numeric values (1 = Yes, 0 = No).
- Replaced empty strings with NaN across all object-type columns.
- Exported cleaned dataset for further analysis.

Task 4: Time Schedule

For our project, we are following this plan:

23.11 – Data controlling and cleaning (4h each)

30.11 – Project report (2h each)

03.12 – Starting to train model(s) and finding correlations between variables (5h each)

04.12-07.12 – Working with our project and reporting our findings (20h each)

07.12 – Last look on our project and submitting (2h each)

We are working on every step together and actively discussing with each ohter, if we come up with new methods to reach our goals. Bigger part of this project will be done in first week of December when we are training our model, clustering and doing analyzis on found correlations.