Capstone 3 Project Report

Mental Health Risk in Tech Workers Post-Pandemic

Prepared by: Maria A. Suarez Springboard

Data Science Career Track

Date: September 2025

1. Introduction

Problem statement: Understanding factors that influence disclosure of mental health conditions among tech workers post-pandemic. This work identifies key predictors, evaluates disclosure likelihood, and generates actionable personas for HR policy insights.

2. Data Overview

- Dataset: Tech Worker Mental Health Survey (□1259 responses after cleaning).
- Key features: age, gender, company size, stigma_index, leave policy, anonymity, benefits, care options, seek help.
- Target variable: Disclosure (1 = would disclose; 0 = would not disclose).

3. Data Wrangling Process

- Loaded survey.csv and standardized column names.
- Removed duplicates and irrelevant columns.
- Handled missing values: median for numeric, mode for categorical.
- Encoded categorical variables using one-hot encoding. –
- Outliers: reviewed stigma scores; capped extreme values when necessary.
- Output: saved survey_clean.csv.

4. Exploratory Data Analysis (EDA)

- [Placeholder for Figure 1: Target distribution chart] Key relationships observed: Leave policy vs. disclosure.
- [Placeholder for Figure 2] - Anonymity vs. disclosure.
- [Placeholder for Figure 3] - Gender, benefits, and care options also show notable trends. [Placeholder for Figure 4] Correlation heatmap shows stigma_index strongly linked with disclosure.

5. Modeling

- Preprocessing: ColumnTransformer pipeline (scaling numeric, one-hot encoding categorical). Train-test split: 80/20 with stratification.
- Models tested: Logistic Regression, Random Forest. Best model: Logistic Regression - CV ROC-AUC: 0.726 - Test ROC-AUC: 0.806 - F1 score: ~0.75 Important features: stigma_index, leave policy, anonymity. [Placeholder for Table 1: Model performance metrics] [Placeholder for Figure 6: Feature importance plot]

6. Segmentation & Personas

- Cluster analysis generated 3–5 employee personas: 1. High stigma, low support. 2. Medium stigma, partial support. 3. Low stigma, high support. These personas help HR leaders target interventions effectively. [Placeholder for Figure 7: Persona cluster visualization]

7. Business Insights

- Clear leave policies strongly increase disclosure likelihood. - Anonymous channels provide safer spaces for disclosure.
- Mental health benefits and support services mitigate stigma effects.
- Recommendations: - Formalize leave policies. - Expand anonymous reporting channels. - Provide comprehensive mental health benefits.

8. Deliverables

- Co Jupyter notebooks (wrangling, EDA, modeling)
- survey_clean.csv dataset. - Dashboards (Tableau/PostHog).
- GitHub repository with documentation.

9. Conclusion & Next Steps

Findings: Logistic Regression achieved strong predictive performance; key drivers include stigma, leave policy, and anonymity. Limitations: Survey-based data; limited geography. Next steps: longitudinal studies; build predictive HR dashboards for real-time monitoring.