OpenLearn Cohort 1.0

Capstone Project 2025

Mental Wellness Analysis and Support Strategy Design for the Tech Workforce

Objective

To understand the key factors influencing mental health issues among employees in the tech industry and build data-driven solutions for:

- **Classification Task**: Predict whether an individual is likely to seek mental health treatment.
- **Regression Task**: Predict the age of an individual based on personal and workplace attributes, supporting age-targeted intervention design.
- **Unsupervised Task**: Segment tech employees into distinct clusters based on mental health indicators to aid in tailored HR policies.

Dataset Overview

Dataset Source: Mental Health in Tech Survey

- Collected by **OSMI** (Open Sourcing Mental Illness)
- Features include:
  - Demographic details (age, gender, country)
  - Workplace environment (mental health benefits, leave policies)
  - Personal experiences (mental illness, family history)
  - Attitudes towards mental health

Project Case Study

As a Machine Learning Engineer at NeuronInsights Analytics, you've been contracted by a coalition of leading tech companies including CodeLab, QuantumEdge, and SynapseWorks. Alarmed by rising burnout, disengagement, and attrition linked to mental health, the consortium seeks data-driven strategies to proactively identify and support at-risk employees. Your role is to analyze survey data from over 1,500 tech professionals, covering workplace policies, personal mental health history, openness to seeking help, and perceived employer support.

Key questions drive your solution:

- Who is most likely to suffer silently and avoid seeking treatment?
- How do factors like remote work, mental health benefits, and managerial support influence mental well-being?
- Can employee profiles be segmented to enable targeted outreach and HR intervention?

OpenLearn Cohort 1.0

Capstone Project 2025

You clean and model the data using advanced classification algorithms to predict likelihood of seeking treatment, and regression models to analyze age-based behavioral trends. Unsupervised clustering uncovers hidden personas such as Silent Strugglers and Supported Advocates helping HR teams tailor programs. Finally, you deploy an interactive Streamlit dashboard enabling HR leaders to explore insights, simulate interventions, and flag high-risk segments anonymously.
Your work bridges the gap between machine learning and mental wellness building not just models, but actionable empathy in the tech workplace.

Project Components

Part 1: Exploratory Data Analysis (EDA)

- Clean missing values and invalid data (especially age and gender anomalies)
- Perform univariate, bivariate, and multivariate analysis
- Visualize distributions, correlations, and trends using matplotlib/seaborn/plotly
- Highlight key patterns: e.g., how company policies affect treatment-seeking behavior

Part 2: Supervised Learning Tasks

A. Classification Task:

- **Objective**: Predict whether a person is likely to seek mental health treatment (treatment column: yes/no)
- **Input Features**: Workplace support, past diagnosis, family history, leave policies, etc.
- **Algorithms to Try**: Logistic Regression, Random Forest, XGBoost, SVM, etc.
- **Evaluation**: Accuracy, ROC-AUC, Confusion Matrix, F1 Score

B. Regression Task:

- **Objective**: Predict the respondent's age
- **Use Case**: Identifying age-based risk factors or bias in workplace culture
- **Input Features**: Workplace type, gender, openness about mental health, remote work, etc.
- **Evaluation**: RMSE, MAE, R^2 score

Part 3: Unsupervised Learning Task

- **Objective**: Cluster tech workers into mental health personas
- **Features to Use**: Aggregated indicators like:
  - Perceived mental health support
  - Openness to discussing mental health
  - Prior experience with mental health issues
- **Algorithms to Try**: KMeans, Agglomerative Clustering, DBSCAN
- **Evaluation**: Silhouette Score, Cluster interpretation
- **Deliverable**: Describe each cluster in layperson’s terms (e.g., "Silent Sufferers", "Open Advocates", "Under-Supported Professionals")

OpenLearn Cohort 1.0

Capstone Project 2025

Part 4: Streamlit Deployment

Create an interactive web app with:

- EDA visualizations
- Input form to predict whether a person will seek treatment
- Display of predicted age based on mental health and workplace features
- Cluster visualizer with description of user persona
- Include model confidence, data summary, and recommendations

Deliverables

1. Technical Report

To include the following structured sections:

- Executive Summary
- Problem Statement
- Dataset Description & Cleaning Summary
- EDA with Visuals
- Classification Task: Modeling, results, and discussion
- Regression Task: Modeling, results, and discussion
- Clustering Analysis & Interpretation
- Deployment Details
- Business Recommendations
- Challenges Faced & Future Scope

2. Streamlit Web App

- GitHub link + hosted Streamlit app (optional)
- Should include all interactive elements for user testing

3. GitHub Repository

- README.md with:
  - Project summary
  - Setup instructions
  - Feature description
  - Link to the Streamlit app deployed
- Code files: EDA.ipynb, classification_model.ipynb, regression_model.ipynb, clustering.ipynb, app.py, requirements.txt, etc.

OpenLearn Cohort 1.0

Capstone Project 2025

Day	Tasks
1	Data Cleaning + EDA
2	Classification + Regression Modeling
3	Clustering + Visualization
4	Streamlit Deployment + Report Writing
5 to 7	Buffer

Suggested Timeline

Evaluation Criteria 100 Marks

Component	Marks
EDA + Visual Insights	15
Classification Task	15
Regression Task	15
Unsupervised Learning Task	15
Model Evaluation + Tuning	10
Streamlit Deployment	10
Report Quality + Structure	10
Innovation / Depth of Work	10