DSA2040A Data Mining Group Project

Group Name

To be updated

Member Names

- Nicholas Kinyanjui - 670178
- Name2 (ID: xxx)
- Name3 (ID: xxx)
- Name4 (ID: xxx)

Project Summary

The OSMI Mental Health in Tech Survey (2014) dataset captures self-reported information from individuals working in the technology sector regarding their experiences, perceptions, and attitudes toward mental health in the workplace. It includes demographic details such as age, gender, country, and employment status, as well as responses about mental health history, access to support resources, workplace culture, and comfort in discussing mental health with employers or colleagues. The dataset aims to highlight the prevalence of mental health issues in the tech industry, identify potential stigma or barriers to seeking help, and promote awareness around mental well-being in professional environments. Collected anonymously, this data provides valuable insights for organizations, researchers, and policymakers interested in improving mental health support and policies in tech-related fields.

ETL Summary

- Raw data is preprocessed to remove formatting issues (extra quotes, semicolons, inconsistent columns).
- Cleaned data is loaded into pandas for further cleaning: handling missing values, standardizing categories, removing duplicates, and engineering new features (e.g., age groups, region).
- The final cleaned dataset is saved for analysis and mining.

Planned Techniques

- **Statistical Analysis:** Descriptive statistics, correlation analysis, group comparisons.
- **Data Mining:**
    - Clustering (e.g., k-means)
    - Classification (e.g., decision trees, logistic regression)
    - Association rule mining (if applicable)
- **Visualization:** Distributions, heatmaps, dashboards (e.g., with Seaborn, Plotly, or Dash)

Tools Used

- Python (pandas, numpy, scikit-learn, matplotlib, seaborn, plotly)
- Jupyter Notebook

- Git & GitHub

Instructions to Run Notebooks

1. Clone the repository:

```
git clone (https://github.com/Testertesting-
create/DSA2040A_DataMining_group1/)
```

2. Install dependencies:

```
pip install -r requirements.txt
```

3. Run the notebooks in order:
   - `notebooks/1_extract_transform.ipynb` (ETL & cleaning)
   - `notebooks/2_exploratory_analysis.ipynb` (EDA & stats)
   - `notebooks/3_data_mining.ipynb` (mining techniques)
   - `notebooks/4_insights_dashboard.ipynb` (dashboard/visualization)

Update this table as the project progresses to reflect actual contributions.

Click the link below to view the TODO list:

👉 Go to TODO list