# Post-Disaster Mental Health Forecasting Using Machine Learning Models

Kanishka Katragadda (A18021511), Sankalp Kumaraswamy(A17282917), Girma Terfa(A18053343)
CSE151A
University of California, San Diego
Kanishka Katragadda ,skumaraswamy@ucsd.edu, gterfa@ucsd.edu

**ABSTRACT**

For this paper, we created machine learning models to predict fear and anxiety following seismic activity in the Campi Flegrei area.

## 1. INTRODUCTION

The issue of post-seismic stress is a unique and interesting topic that can bring light to issues faced by survivors of disasters around the world. Performing exploratory data analysis can provide insights into this issue and building predictive models can assist with providing assistance and intervention for those struggling psychologically following seismic activity. Projects like these represent contributions to the understanding of earthquake-induced stress and provide a foundation for further research on mitigating the impact on exposed populations.

## 2. METHODS

### DATA EXPLORATION

The dataset contains 472 observations and 43 features, a mix of categorical and numerical variables. Categorical variables include features like sex, marital status, salary, political orientation, and employment regions, while numerical features are standardized on scales from 0 to 5 or presented as ordinal values such as salary ranges (e.g., "Below €15000" or "Between €15000 and €28000"). A significant portion of the data is missing for critical features, with 36% null values in salary and 33% in political orientation. Target variables, including fear, anxiety,

insomnia, and physiological symptoms, are directly tied to seismic stress and were analyzed for correlations. Correlation analysis revealed multicollinearity issues, particularly between fear and anxiety, prompting the selection of a single target variable (fear) to avoid redundancy. The distribution of gender, age, and other features was reviewed to identify potential imbalances or patterns requiring preprocessing adjustments.

**DATA PREPROCESSING**

Preprocessing was conducted systematically to prepare the dataset for modeling. Missing values were addressed using a combination of removal and imputation. For features such as sex, age, and family disabilities, null values were deemed insignificant in volume and were removed entirely. For features with a higher proportion of missing data, such as salary and political orientation, imputation strategies were applied—using the mode for salary and substituting "None" for political orientation, which was already a valid category. Categorical variables required encoding to make them usable for modeling. One-hot encoding was employed for the linear regression model to generate binary columns for each category, ensuring compatibility with the algorithm. For the decision tree model, label encoding replaced one-hot encoding, reducing dimensionality and simplifying data representation. Numerical features, already standardized, were left unchanged. Dimensionality reduction was also performed by removing features with weak correlations to the target variable (correlation $<0.15$) based on the correlation matrix. This step ensured that the dataset focused on meaningful predictors and reduced noise.

**MODEL 1**

For the linear regression model, the dataset was prepared using one-hot encoding for categorical variables. This significantly increased the number of features, adding binary columns for each

category. To address multicollinearity features with correlations below 0.15 with the target variable were excluded from the dataset. Missing values were handled by removing rows with nulls where feasible, particularly for binary or categorical features, as imputing them would introduce noise. The data was then split into training (80%) and testing (20%) sets. Linear regression was selected as a baseline model to capture linear relationships between the features and the target variable. Regularization techniques, such as Ridge regression, were explored to address potential overfitting. The model was implemented with these preprocessing steps to evaluate its ability to generalize the data and provide insights into the strengths and weaknesses of a linear approach.

**MODEL 2**

The decision tree regression model used a preprocessing pipeline optimized for handling categorical data natively. Categorical variables were encoded using label encoding instead of one-hot encoding to streamline the dataset and reduce dimensionality. This approach allowed the decision tree to directly split on categories without creating additional columns. Handling missing values was tailored to the dataset's needs. Rows with null values in features such as sex and age were removed due to their minimal impact on the overall dataset. For salary and political orientation, missing values were replaced with the mode and "None," respectively, as these values provided meaningful and representative imputation. Hyperparameter tuning was conducted to optimize the decision tree structure, with maximum depth ranging from 1 to 20. This process helped balance the complexity of the model and mitigate overfitting by determining the optimal pruning level. The processed data was split into training (80%) and testing (20%) sets. The decision tree structure was visualized to review its splitting patterns and ensure the

model made logical decisions. These steps tailored the decision tree regression to the dataset's characteristics, enhancing its ability to capture non-linear relationships and patterns.

## 3. RESULTS

The preprocessing steps effectively prepared the dataset for modeling, addressing missing values and reducing dimensionality while ensuring compatibility with the chosen algorithms. Since the predictions were multi-output, we had to clip our outputs to specific labels. For the linear regression model, one-hot encoding significantly increased the number of features, leading to an R-squared score of 0.78 on the training set but a test Mean Squared Error (MSE) of 0.5259, indicating overfitting. The inclusion of Ridge regularization marginally improved generalization but did not fully mitigate the overfitting problem. In contrast, the decision tree regression model achieved a more balanced performance. After label encoding reduced the dataset's dimensionality, hyperparameter tuning optimized the tree's maximum depth, resulting in a training MSE of 0.3366 and a test MSE of 0.3734. This minimal gap between training and testing performance demonstrated the model's ability to generalize effectively. Tree visualization confirmed logical and interpretable splits, showcasing the model's strength in capturing non-linear relationships and handling categorical data more efficiently than the linear regression approach. Overall, the decision tree model outperformed the linear regression model in terms of generalization and predictive accuracy.

## 4. DISCUSSION

The process of building machine learning models for post-disaster mental health forecasting

revealed both promising outcomes and critical challenges. Each phase of the project highlighted areas where careful decisions were needed, from data handling to model evaluation.

**Preprocessing**: The preprocessing phase underscored the importance of tailoring techniques to the dataset's specific characteristics. By opting for imputation strategies aligned with the nature of each feature—such as replacing null salary values with the mode and encoding categorical variables differently for each model—we aimed to balance data integrity and usability. However, these decisions were not without trade-offs, as imputation inherently risks oversimplifying complex patterns and dropping rows, even a small percentage, could have introduced subtle biases. These trade-offs, though necessary, illustrate the nuanced challenges of working with real-world datasets.

**Model Development:** The shift from linear regression to decision trees reflected a broader need to align modeling approaches with data complexity. Linear regression, while a solid baseline, struggled to capture the intricacies of non-linear relationships in the data, as evidenced by its overfitting tendencies. Decision trees, with their ability to split on categorical data natively, addressed this limitation effectively. The pruning process during hyperparameter tuning proved particularly valuable, allowing for a balance between model interpretability and performance.

**Results and Insights:** The performance metrics indicate that decision trees provided a more nuanced understanding of the target variable compared to the linear model. Visualization of the tree's structure offered an additional layer of interpretability, highlighting logical decision splits that could inform further refinements. Still, while the results are promising, the relatively narrow focus on fear as a singular target variable leaves room to explore the broader psychological impacts of seismic events, including potential interactions between fear and other related symptoms.

**Possible Limitations:** Throughout the project, several limitations became apparent. The dataset's inherent constraints, including missing values and its relatively small size, posed challenges to model generalizability. Moreover, while decision trees performed well, they are inherently prone to overfitting in deeper configurations, which may limit their utility in datasets with higher dimensionality or noisier features. These constraints highlight the need for ongoing refinement and exploration of more advanced techniques in future work.

## 5. CONCLUSION

Reflecting on this project has led to some clear insights about this project and dataset that we did not know going into. A major struggle we encountered was the lack of predictive power the majority of our features gave us. While our decision tree regressor was able to achieve good performance, it was reliant on the strong relationship between fear and anxiety and did most of the splits on those features. We attempted to build a model without the anxiety feature to predict fear and it massively increased our loss and reduced our R-squared score. It's possible that a higher complexity model like ANNs would have been able to overcome this, but again there is the issue of overfitting. There was some supplementary information gain from housing type, shocks, and decision timeliness, but most of our features had zero importance to the model. This calls into question the nature of the data itself. Should we have tried to create new features? How would we have done that? Does the dataset inherently lack the power to predict fear and anxiety as a single feature? While quite comprehensive, the majority of the features within the dataset have little if any relevance to seismic stress. For future research in this area, it may be better advised to look at a smaller number of more relevant features such as psychological counseling rather than having many features dedicated to government and institutional trust that ended up having no impact statistically. Longitudinal data would also be helpful in creating time-series models that may better capture the long-term effects of seismic stress. These are just some of the

options to take into consideration when continuing research in this area of study to improve the lives of those impacted by these life-altering disasters.

## 6. STATEMENT OF COLLABORATION

Kanishka Katragadda: Title: Coding, and meeting setup
Sankalp Kumaraswamy : Title: Coding and writing.
Girma Terfa: Title: Coding and Submissions.