CPSC 6185: Final Project

KHUSHI JANI

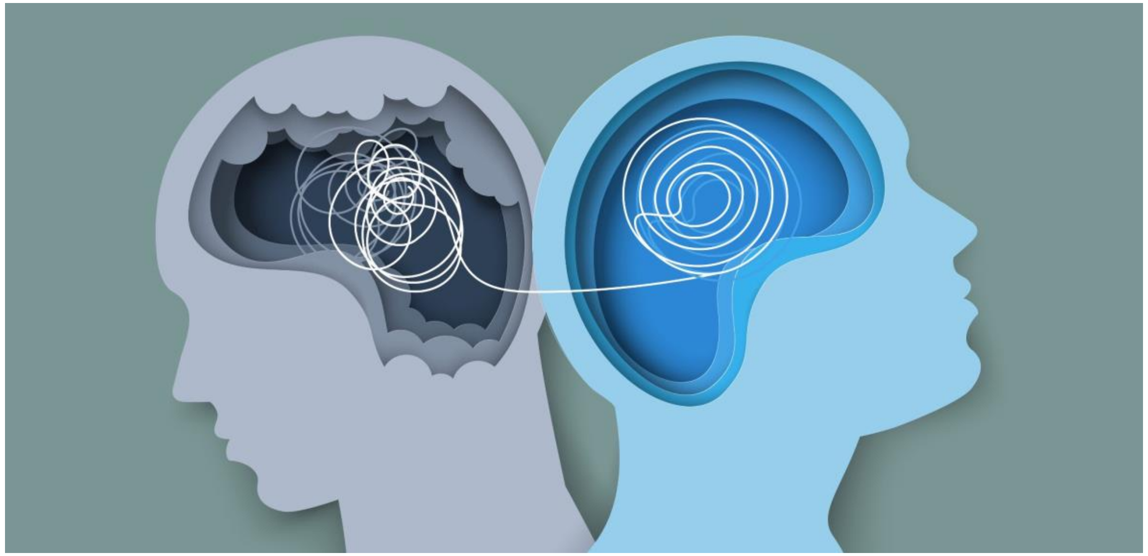
KORIE MACDOUGALL

VENKAT RAMANA REDDY KUPPI REDDY

Project Overview and Goals

- Goal
  - To classify and predict treatment attribute (whether the individual sought mental health treatment) based on the construction of a decision tree and comparison against clusters of the original dataset
- AI Technique(s)
  - Decision tree
    - Target feature: treatment
  - K-means clustering
- **Dataset**: Mental Health in Tech Survey | EDA
  - https://www.kaggle.com/code/chaitanya99/mental-health-in-tech-survey-eda/input
  - 27 features reflecting attitudes towards mental health and frequency of mental health disorders
  - 1259 rows, 27 columns
- **Domain**: Mental health in tech field

Feature Selection by Intuition



- **Age** – Respondent age
- **Gender** – Respondent gender
- **Self-Employed** – Are you self-employed?
- **Family History** – Do you have a family history of mental illness?
- **Remote Work** – Do you work remotely (outside of an office) at least 50% of the time?
- **Tech Company** – Is your employer primarily a tech company/organization?
- **Benefits** – Does your employer provide mental health benefits?
- **Care Options** – Do you know the options for mental health care your employer provides?
- **Seek Help** – Does your employer provide resources to learn more about mental health issues and how to seek help?
- **Observed Consequences** – Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- ***Treatment*** *– Have you sought treatment for a mental health condition?*

Dataset Cleaning and Preparation

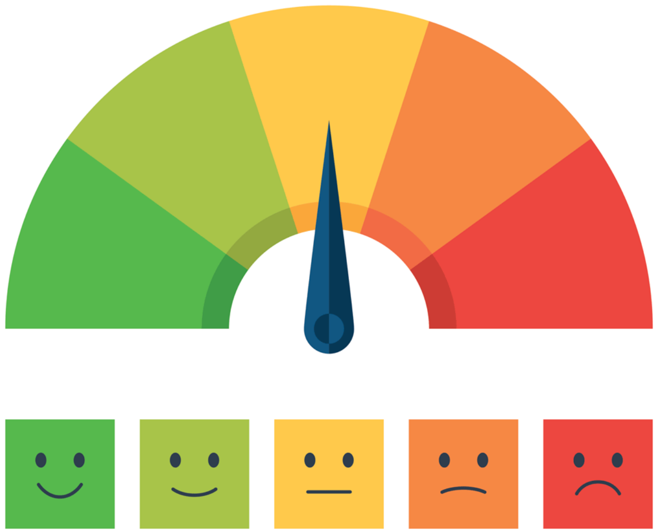
- 4 features had missing values that were replaced to avoid dropping any records and losing data.
  - **state**, **self_employed**, and **work_interfere** were filled with "Unknown"
  - **comments** was filled with empty strings
- Columns with inconsistent values/formatting were standardized to ensure uniformity and improve interpretability
  - **Age** had some negative/unrealistic values, so values were limited to be 18-120 otherwise NaN to avoid skewing the data.
  - **Gender** had an array of open response values, so responses were mapped to their category: Female, Male, or Other to allow meaningful analysis.
  - Remaining features already used standardized categorical values, so no changes were made.

Encoding the Dataset

- Decision trees split based on numeric thresholds, so categorical inputs were converted to numerical form.
- One-hot encoding was applied to nominal categorical features (**gender**, **country**, **state**, **self_employed**, etc), which allows models to treat each category as a separate binary feature.
- Ordinal categorical features (**work_interfere**, **no_employees**, and **leave**) were mapped to numeric values.
- **comments** was converted to a binary column indicating if a comment was provided.

Scaling the Dataset

- K-means clustering uses Euclidean distance to measure distance between data points, which can be skewed if features are on different scales.
- Z-score standardization was used to scale features to have a mean of 0 and a standard deviation of 1 to ensure all features contribute equally to the clustering.



Oversampling Data

- Our original dataset was already balanced, with nearly equal numbers of records labeled as **treatment**=no and **treatment**=yes.
- After data cleaning and standardizing feature values, several features had excessive records with "Unknown" values.
- For the decision tree model, records with "Unknown" values were removed and **treatment**=no records were oversampled to maintain balance and improve model performance.

Decision Tree Model

- Used sklearn to build decision trees
- Trained the tree on the oversampled, hot-encoded data
- Fine-tuned model via GridSearchCV
  - Evaluated combinations of parameters through cross-validation
    - Max_depth – limits tree depth
    - Min_samples_split – minimum samples to split a node
    - Criterion – gini and entropy
  - Selected the best model based on accuracy scoring metric
  - The GridSearchCV may be overfitting the model on the full dataset because accuracy decreased

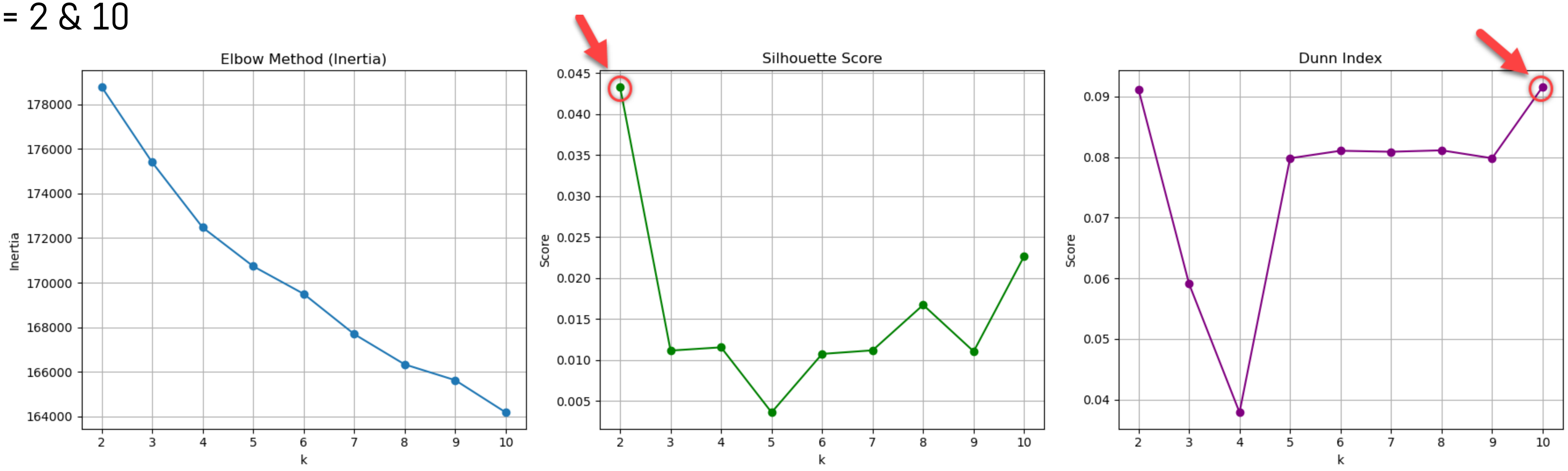
Dataset	Accuracy Before	Max_depth	Min_samples_split	Criterion	Accuracy After
Intuitive Selection	0.8	10	2	entropy	0.8
Full	0.897	10	2	entropy	0.795

K-Means Clustering

- Used sklearn for clusters on hot-encoded, stratified data including 'Unknown' values
- Defined the function dunn_index to evaluate clustering quality for different values of k (2-10)
- Tracked 3 metrics for optimum k:
  - **Inertia** – measures how internally coherent the clusters are (lower = tighter clusters)
  - **Silhoutte score** – measures how well each point fits in its cluster (closer to +1 = distinct clusters)
  - **Dunn Index** – measures cluster compactness and separation (higher = far apart and compact)

Intuitive Selection: k = 3 & 10

Full: k = 2 & 10



Decision Tree Results

- Evaluation Metrics
  - Accuracy – the proportion of correctly classified instances among all instances
    - Overall effectiveness
  - Precision – the proportion of true positive predictions out of all positive predictions
    - Understanding the likelihood of a false positive
  - Recall – the proportion of true positive predictions out of all positive instances
    - Understanding the likelihood of false negatives
  - F1-Score – single measure of performance based on both precision and recall equally

Prediction	Dataset	Accuracy	Precision	Recall	F1-Score
No	Intuitive Selection	0.80	0.77	0.85	0.81
	Full	0.897	0.90	0.90	0.90
Yes	Intuitive Selection	0.80	0.83	0.75	0.79
	Full	0.897	0.89	0.89	0.89

Confusion Matrix

The Full dataset performed better overall

Intuitive Selection

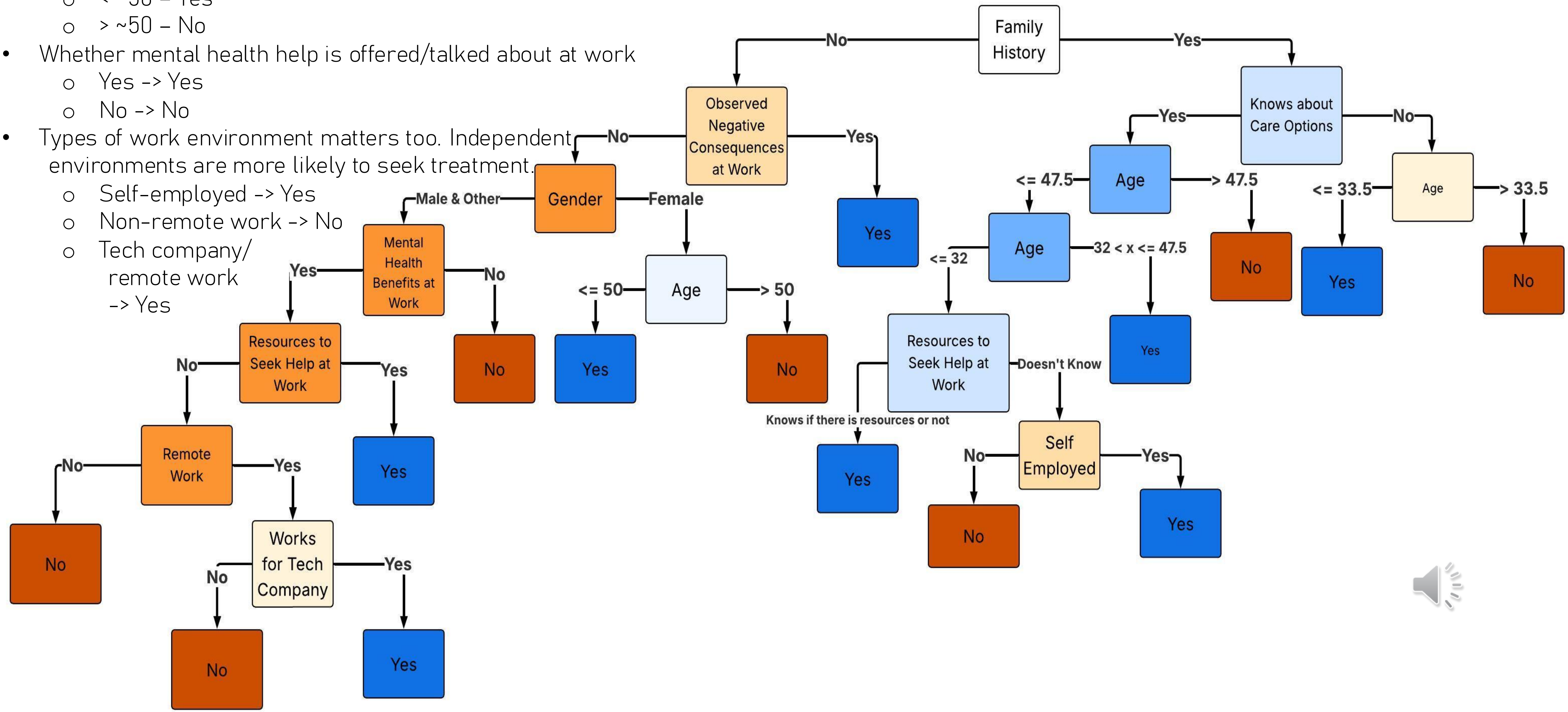
	Predicted Positive	Predicted Negative
Actual Positive	17	3
Actual Negative	5	15

Full

	Predicted Positive	Predicted Negative
Actual Positive	18	2
Actual Negative	2	17

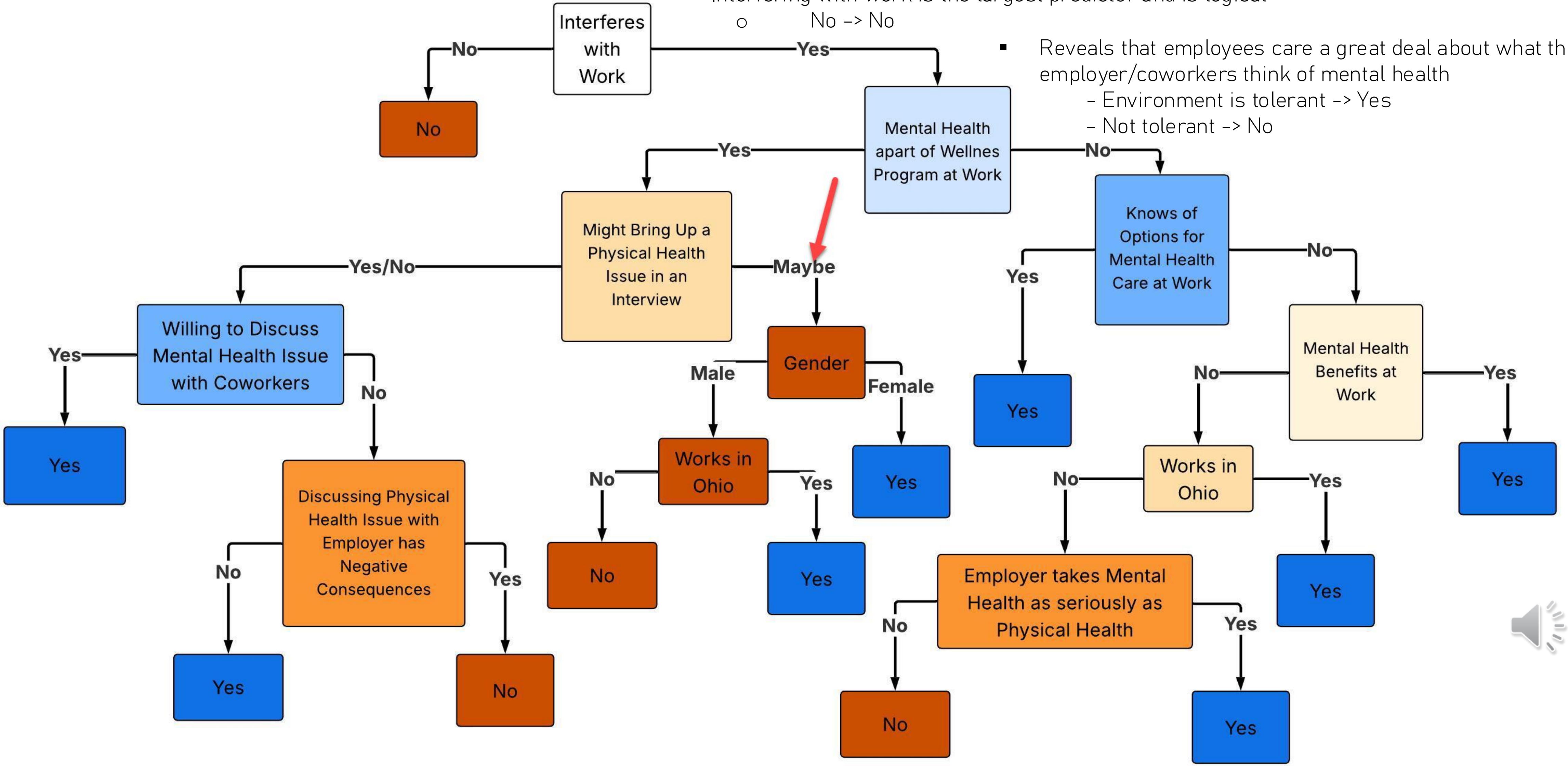
Intuitive Selection

- Age is strong predictor
  - < ~50 – Yes
  - > ~50 – No
- Whether mental health help is offered/talked about at work
  - Yes -> Yes
  - No -> No
- Types of work environment matters too. Independent environments are more likely to seek treatment.
  - Self-employed -> Yes
  - Non-remote work -> No
  - Tech company/ remote work -> Yes



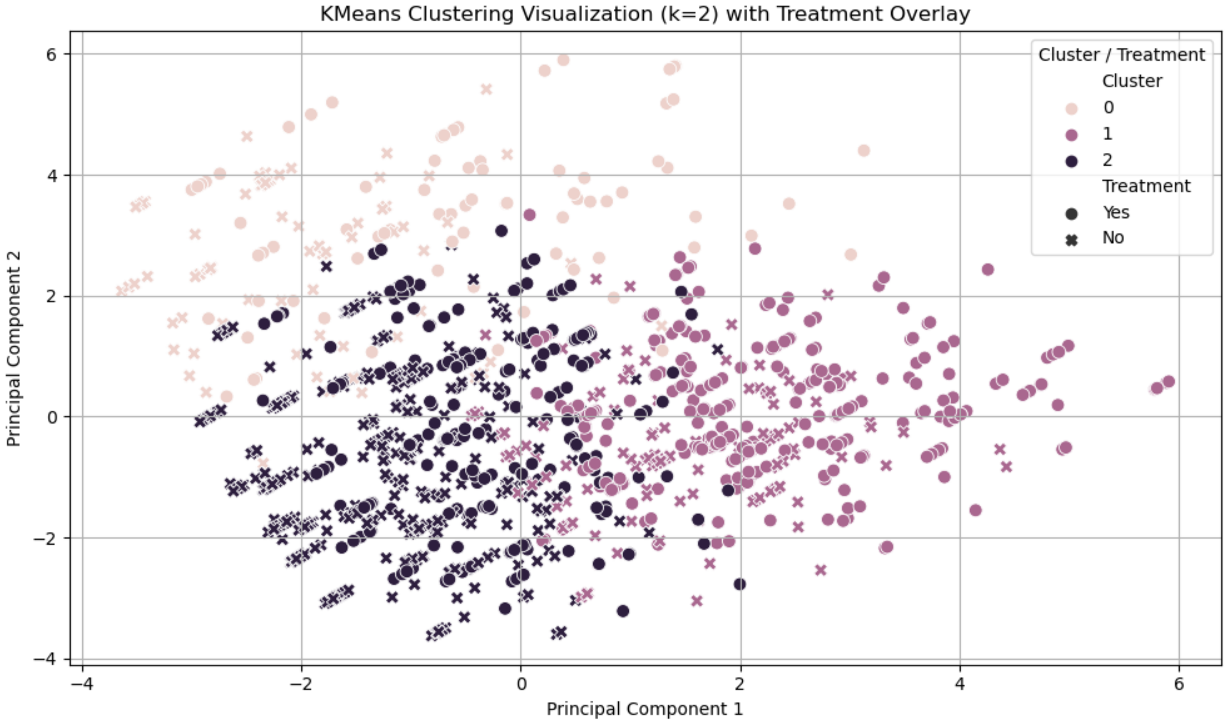
Full Dataset

- Reflects selective dataset in that gender is also a strong predictor
  - Female - Yes
- Data may be too noisy with uncertain answers like 'Maybe' and 'Don't Know' that don't define a clear path
- Working in the state of Ohio is an unusual, yet strong predictor
  - Yes -> Yes
  - Interesting topic to explore further – what about working in Ohio helps people seek treatment?
    - Interfering with work is the largest predictor and is logical
      - No -> No
        - Reveals that employees care a great deal about what their employer/coworkers think of mental health
          - Environment is tolerant -> Yes
          - Not tolerant -> No



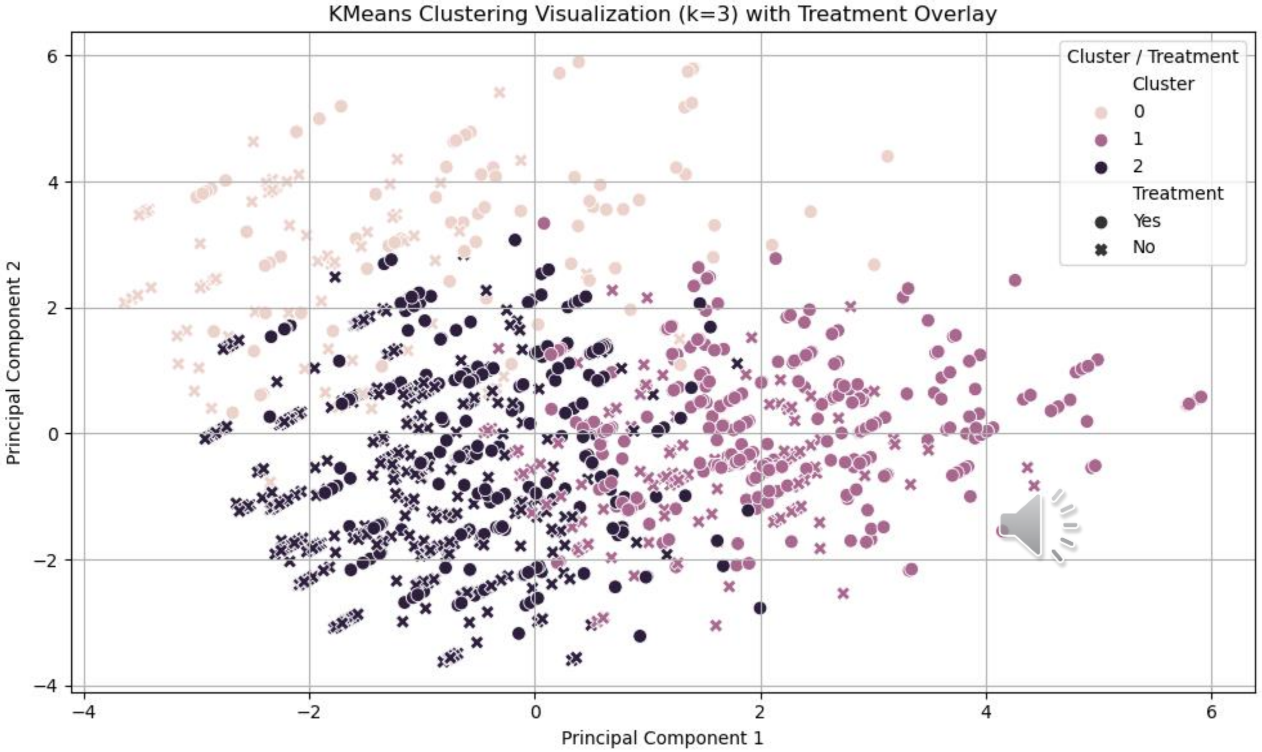
K-Means Clustering Results

Full



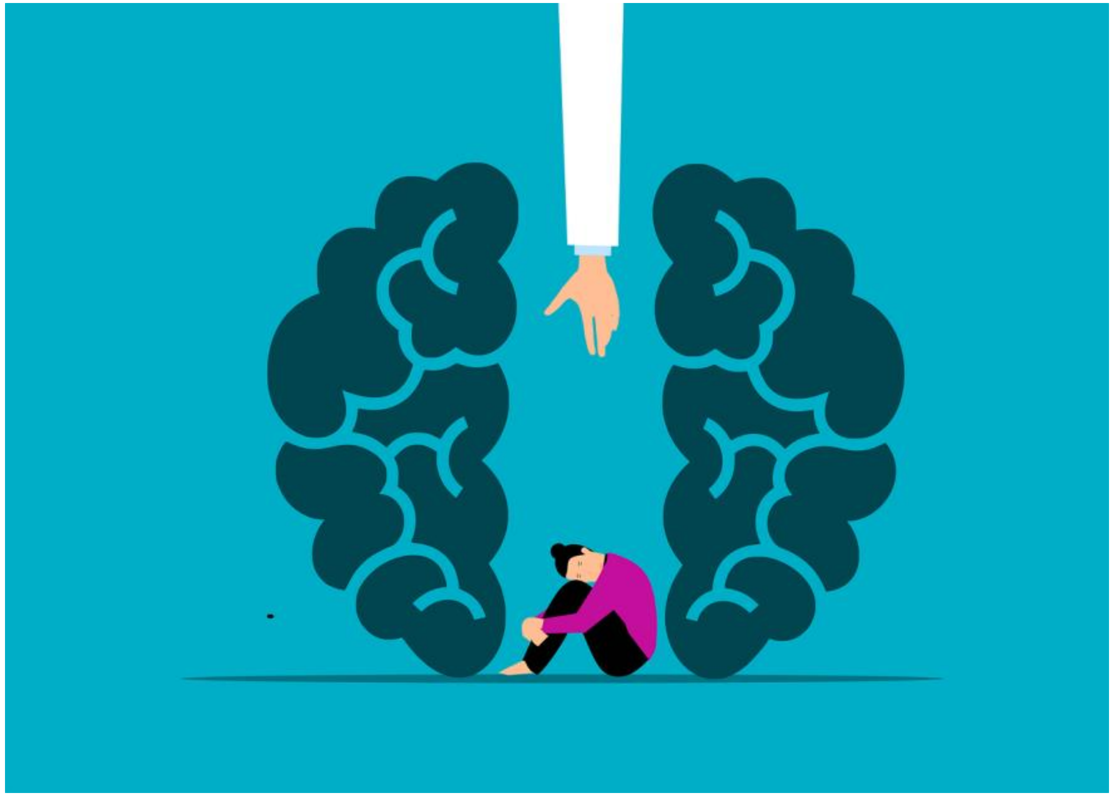
Dataset	Silhouette	Dunn Index
Intuitive Selection	0.14	0.184
Full	0.043	0.092

Intuitive Selection



- Performed poorly overall
- Treatment categories did not form clear and separated clusters
- May be because of too noisy data and Euclidean distance is less compatible with categorical variables
- Smaller, oversampled dataset with no missing values performed just as poorly

Overall Insights



- Decision Tree was a much better model and clearly defined features that could predict whether someone sought treatment or not
  - Main splits on workplace support/resources and attitude/tolerance towards mental health issues
  - Defines a clear generational and gender divide – result that is consistent with current studies by the CDC[1]
    - Gives more legitimacy to other findings
  - Unusual finding on working in Ohio
  - Intuitive selection excluded one of the most predictive features – work interference
  - Employees with more independence and control over their environment more likely to seek help
- Could perform better if more features were included and noisier features with 'Maybe' or uncertain answers were excluded
- Could try using k-modes instead because it is more suitable for categorical variables[2]

[1]Terlizzi EP, Norris T. Mental health treatment among adults: United States, 2020. NCHS Data Brief, no 419. Hyattsville, MD: National Center for Health Statistics. 2021. DOI: https://dx.doi.org/10.15620/cdc:110593.

[2]Masego. (2025, January 24). *Making sense of categorical survey data: A K-Modes clustering case study*. Medium. https://medium.com/@masego_m/making-sense-of-categorical-survey-data-a-k-modes-clustering-case-study-e492684bfc40