1. Problem Definition

Mental health is a critical concern in the technology industry, where employees often face high stress, long working hours, and limited access to mental health resources.

The goal of this project is to analyze the **"Mental Health in Tech Survey"** dataset to:

- Explore patterns and trends related to mental health issues among tech workers.
- Predict whether an employee seeks treatment based on personal and workplace factors.
- Discover underlying clusters of individuals with similar mental health characteristics.

By understanding these patterns, the project aims to support both individual well-being and organizational effectiveness, demonstrating the value of data-driven approaches in addressing mental health challenges in the workplace.

2.Software Development Fundamentals

Purpose:

The purpose of applying software development fundamentals in our project is to ensure that the **data analysis, machine learning models, and reporting workflow** are structured, maintainable, and reproducible.
By following a systematic process, the project achieves **accuracy, scalability, and clarity**, which are critical for the success of data-driven solutions.

Our key followed steps:

Step No.	Development Step	How we use it	Contribution to Success
1	**Problem Definition**	Clearly define the main objective ; analyzing factors influencing mental health in the tech industry.	Keeps the project focused and goal-oriented, preventing wasted effort.
2	**Requirements Analysis**	Identify dataset source (Kaggle: OSMI Survey), tools (Python, Pandas, Scikit-learn), and deliverables (visualizations, report).	Establishes clear expectations, ensuring appropriate resource and tool selection.
3	**Data Collection**	Retrieve the dataset using KaggleHub and confirm data integrity.	Ensures reliable and consistent data sourcing for reproducible analysis.
4	**Data Preprocessing**	Handle missing values, remove inconsistencies, encode categorical	Improves model accuracy and prevents data quality issues.

		variables, and normalize numerical features.	
5	**Model Development**	Build and train machine learning models (Logistic Regression, Random Forest, K-Means, Hierarchical Clustering).	Forms the analytical foundation, providing measurable insights and predictions.
6	**Testing & Validation**	Evaluate models using metrics such as Accuracy, F1-Score, and Silhouette Score.	Confirms the reliability and effectiveness of the models before final reporting.
7	**Visualization & Reporting**	Creating a visual comparisons by (bar charts, Histogram, radar charts and Elbow curve) and summarize findings.	It makes complex data in an understandable and impactful way by seeing it as a graph .
8	**Documentation & Deployment**	Prepare a professional report and ensure all findings and code are documented clearly.	Guarantees reproducibility, clarity, and professionalism in the final solution.

How These Steps Contribute to our Project Success?

- **Clarity:** Each stage provides direction, ensuring consistent progress.
- **Quality Assurance:** Preprocessing and testing enhance the reliability of analytical outcomes.
- **Reproducibility:** Automated data collection and scripted analysis ensure repeatability.
- **Scalability:** The modular design allows additional models or datasets to be integrated later.
- **Professionalism:** Following structured software development practices mirrors real-world data science workflows.

---

3. Dataset Description

Source: our source is from Kaggle – OSMI: Mental Health in Tech Survey
Number of Entries: 1,259
Number of Columns: 27

Sample Columns:

Column	Description
`Age`	Participant's age
`Gender`	Gender identity
`Country`	Country of residence
`self_employed`	Whether self-employed

#	Timestamp	Age	Gender	Country	State	Self-Employed	Family History	Treatment	Work Interfere	No. of Employees
0	2014-08-27 11:29:31	37	Female	United States	IL	—	No	Yes	Often	6–25
1	2014-08-27 11:29:37	44	M	United States	IN	—	No	No	Rarely	More than 1000
2	2014-08-27 11:29:44	32	Male	Canada	—	—	No	No	Rarely	6–25
3	2014-08-27 11:29:46	31	Male	United Kingdom	—	—	Yes	Yes	Often	26–100
4	2014-08-27 11:30:22	31	Male	United States	TX	—	No	No	Never	100–500

Column	Description
`family_history`	Family history of mental illness
`treatment`	Whether the participant sought treatment
`work_interfere`	How often work interferes with mental health

Data Overview

- Missing values found in several columns such as `state`, `work_interfere`, and `comments`.
- Some age outliers (e.g., unrealistic high values).
- Total duplicates: **0**

[Figure Table 1: Sample of Dataset Head]

Figure1 :As we can see in the Sample Head Dataset, there are some missing values.

---

4. Data Preprocessing

4.1 Handling Missing Values

- Removed rows with missing key fields (`Age`, `Gender`, `work_interfere`, `family_history`, `remote_work`).
- Filled missing `self_employed` entries with “No”.
- Imputed missing `Age` values with the median age.

4.2 Feature Selection

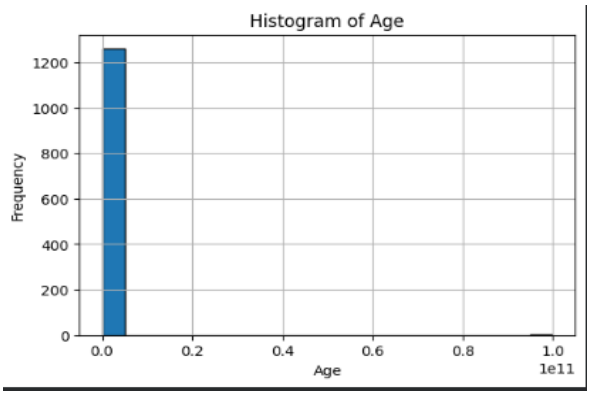
Selected relevant columns for supervised and unsupervised learning:

- **Demographics:** `Age`, `Gender`, `Country`
- **Work factors:** `remote_work`, `family_history`, `self_employed`
- **Target variable:** `treatment` (Yes = 1, No = 0)

4.3 Encoding & Scaling

- Converted categorical variables into numeric form using **Label Encoding** or **One-Hot Encoding**.
- Scaled numerical features using **StandardScaler** to normalize feature ranges.

[Figure 2: Histogram of Age Distribution]



---

5. Supervised Learning Models

5.1 Model 1 – Logistic Regression

- A baseline model suitable for linearly separable data.
- **Training:** 80% of the data
- **Testing:** 20% of the data
- **Performance:**

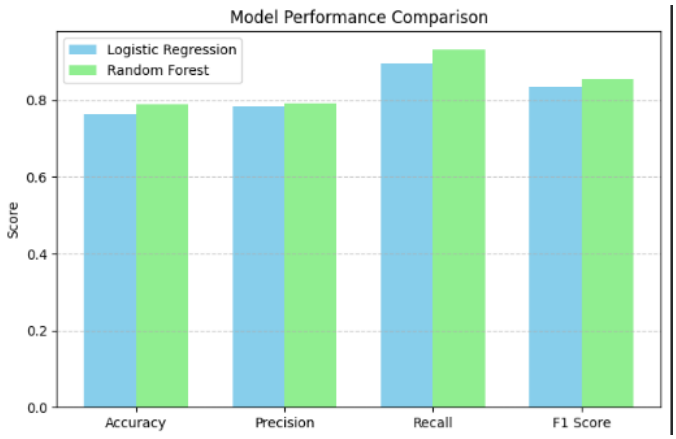
Metric	Score
Accuracy	0.76
Precision	0.70
Recall	0.90
F1-Score	0.84

5.2 Model 2 – Random Forest Classifier

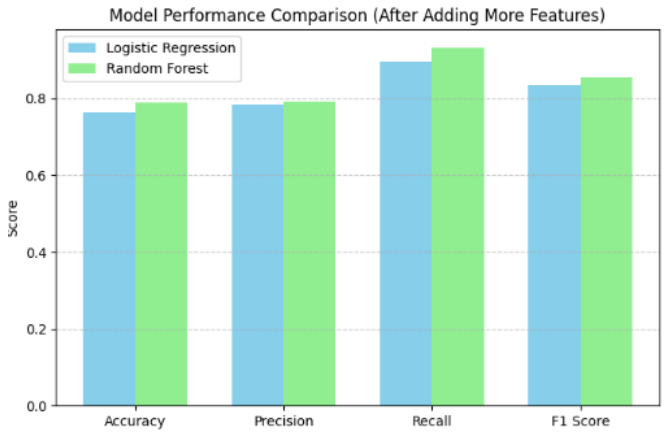
- Ensemble method capable of modeling non-linear relationships.
- **Performance:**

Metric	Score
Accuracy	0.79
Precision	0.79
Recall	0.93
F1-Score	0.86

[Figure 3: Model Performance Comparison Bar Chart]



[Figure 3.1: Model Performance Comparison Bar Chart (After Adding More Features]



5.3 Interpretation

- Random Forest outperformed Logistic Regression in all metrics.
- Indicates that the dataset exhibits **non-linear relationships** between features and the mental health treatment outcome.

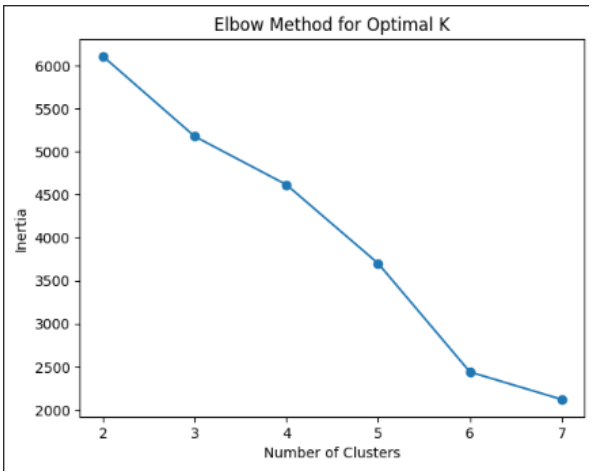
---

6. Unsupervised Learning Models

6.1 K-Means Clustering

- Used the **Elbow Method** to determine optimal cluster count (K ≈ 3).
- **Silhouette Score:** 0.325
- Quick and computationally efficient, but less precise for complex data.

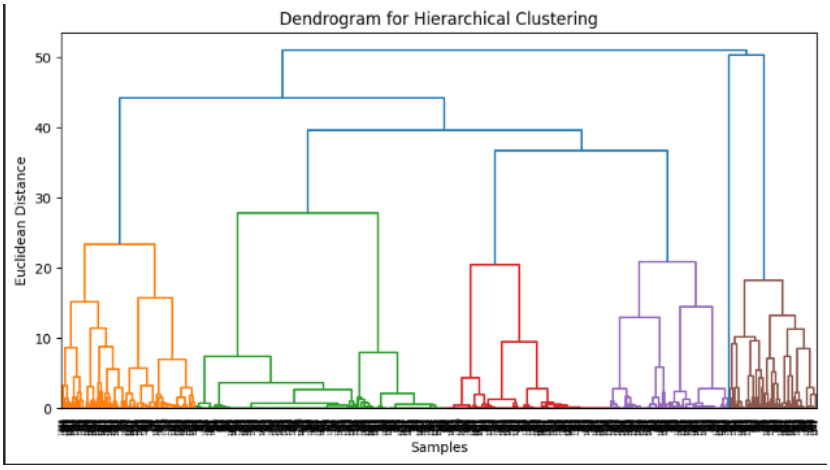
[Figure 4: Elbow Curve for K Selection]



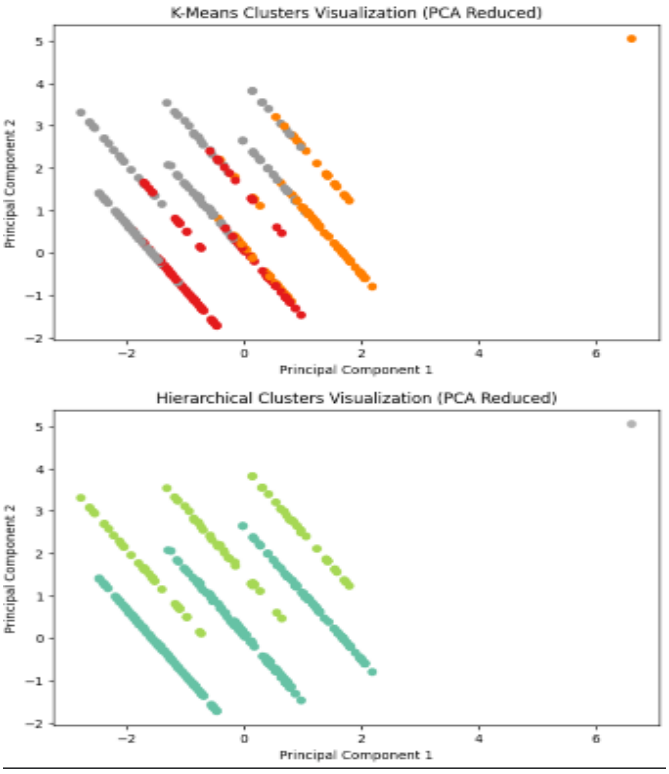
6.2 Hierarchical Clustering

- Revealed a clearer structure in the dataset.
- **Silhouette Score:** 0.35 (higher than K-Means).
- Suggests Hierarchical Clustering better captured group patterns
-

[Figure 5: Dendrogram Plot]



[Figure 6: PCA Visualization of Hierarchical Clusters]

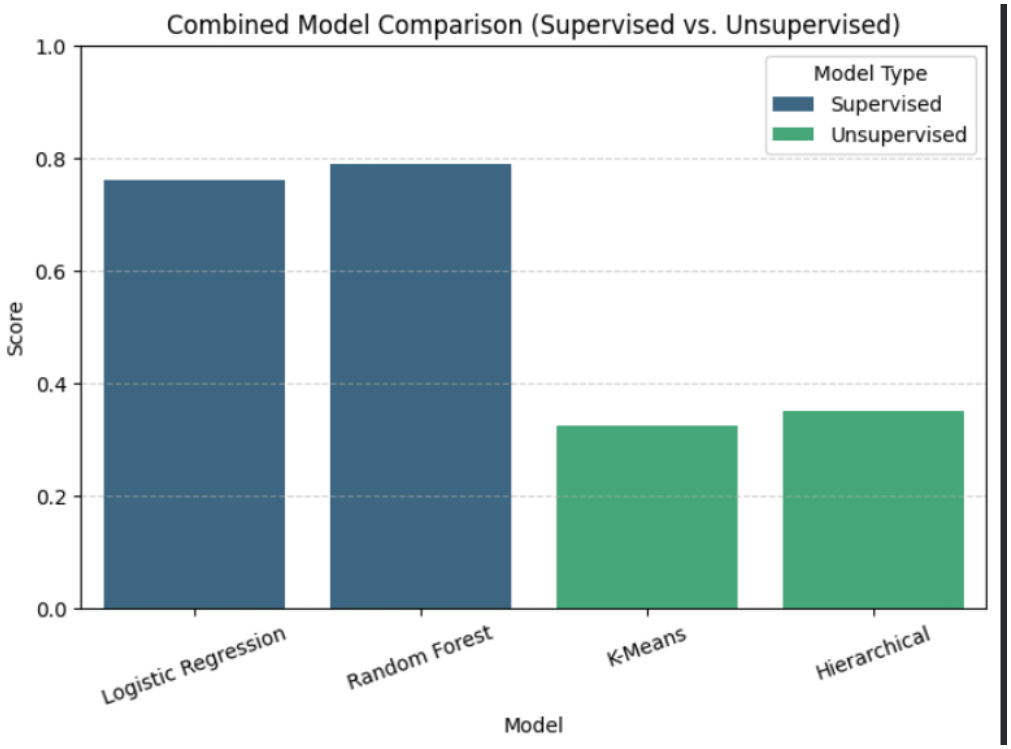


---

7. Model Comparison Summary

Category	Model	Key Strength	Accuracy / Score
Supervised	Logistic Regression	Simplicity, interpretability	76% accuracy
Supervised	Random Forest	Handles non-linearity, higher performance	**79% accuracy**
Unsupervised	K-Means	Fast & scalable	Silhouette: 0.325
Unsupervised	Hierarchical	Captures nested relationships	**Silhouette: 0.35**

[Figure 7: Combined Model Comparison Visualization]



---

8. Challenges Encountered:

- **Data Quality Issues:** Several missing values and unrealistic age entries.
- **Imbalanced Responses:** Some categorical levels were underrepresented.
- **Textual Variations:** Gender column required normalization (e.g., “male”, “**M**”, “man”).
- **Complex Relationships:** Required tree-based methods for better performance.

---

9. Conclusions

- **Random Forest** achieved the best supervised learning performance (≈79% accuracy).
- **Hierarchical Clustering** provided better unsupervised grouping (Silhouette 0.35).
- Mental health in tech appears strongly influenced by **family history**, **work interference**, and **company culture** (remote work, benefits, etc.).
- Encourages organizations to provide **accessible mental health support** and **open discussion environments** to reduce stigma.

10. References

- OSMI Mental Health in Tech Survey (Kaggle)
- Scikit-learn Documentation (v1.4+)
- Pandas, NumPy, and Matplotlib libraries