**Project Documentation: Early Suicide Prediction Using Machine Learning and Data Analysis**

**1. Project Overview**

This project addresses the critical and sensitive issue of mental health and suicide attempts. Its primary focus is to leverage data analysis and machine learning techniques to identify key factors that may lead to suicidal ideation or attempts. By building a predictive model, the project aims to contribute to the early identification of at-risk individuals, enabling timely intervention and support.

**1.1. Project Motivation (The Data Story)**

Suicide is a deeply important and urgent issue that affects millions of people worldwide. The team chose to focus on suicide risk prediction because of its potential to save lives by enabling early intervention for those who need it most.

Many existing approaches to suicide prevention rely heavily on self-reporting or clinical judgment, which can sometimes miss individuals who are silently struggling. By leveraging data-driven methods, this project aims to provide a more objective, scalable, and timely way to identify high-risk individuals. Furthermore, suicide prevention is not only a medical or psychological challenge but also a social responsibility. By addressing this topic, we aim to contribute to the broader effort of reducing stigma around mental health and encouraging open conversations.

Ultimately, this topic was chosen based on the belief that technology and data science can be powerful tools to make a real, positive impact on people’s lives and to help prevent the tragic loss of life caused by suicide.

**1.2. Project Purpose**

To analyze and clarify the psychological and social factors associated with the risk of suicide, thereby enabling mental health professionals to better monitor at-risk individuals and make proactive, data-informed decisions.

**2. Project Objectives**

1. **Data Analysis:** To understand the patterns and correlations between various psycho-social factors (e.g., age, stress levels, mental support) and an individual's propensity for suicidal attempts.
2. **Predictive Modeling:** To develop and evaluate a machine learning model capable of classifying individuals based on their suicide attempt risk level (Never Thought, Thought, Attempted).
3. **Data Visualization:** To create an interactive Power BI dashboard that presents the analytical findings in a clear, accessible, and insightful manner for stakeholders.

**3. Team Members and Roles**

The project workload was efficiently distributed among the team members to cover all project phases:

* **Mohamed Yahya:**
  + **Role:** Data Preprocessing.
  + **Responsibilities:** Data cleaning, handling missing values, outlier detection, and preparing the dataset for modeling and analysis.
* **Ahmed Zakaria Hamed & Mahmoud Reda Bassuony:**
  + **Role:** Model Building.
  + **Responsibilities:** Implementing various machine learning algorithms, training and testing the models, evaluating their performance, and fine-tuning them for optimal results.
* **Basma Nabil Deniour:**
  + **Role:** Power BI Analysis & Visualization.
  + **Responsibilities:** Data cleaning in Power Query, dashboard design, creating the main Overview Page, and developing the Mental Health Distribution analysis page.
* **Yasser Shawky Kashef Ghonem:**
  + **Role:** Power BI Analysis & Visualization.
  + **Responsibilities:** Data modeling, dashboard design, contributing to the Overview Page, and developing the Suicide Behavior Analysis page.

**4. Methodology and Workflow**

The project followed a structured workflow with two parallel tracks: a Machine Learning track and a Business Intelligence (BI) track.

1. **Data Acquisition:** The project utilized the Early Suicide Prediction.csv dataset.
2. **Machine Learning Track:**
   * **Data Preprocessing:** Analyzing, cleaning, and transforming the data using Python.
   * **Feature Engineering:** Converting categorical features into a numerical format using One-Hot Encoding.
   * **Model Building:** Training and evaluating multiple models, including Decision Tree, Random Forest, and a Neural Network.
   * **Optimization:** Addressing the class imbalance problem and fine-tuning model hyperparameters.
3. **Business Intelligence (BI) Track:**
   * **Data Cleaning & Modeling:** Using Power Query for data cleaning and establishing relationships between data fields (Data Modeling).
   * **Dashboard Development:** Designing interactive pages to visualize key statistics and insights based on a predefined set of analytical questions.

**5. Technical Phase: Data Analysis & Preprocessing**

This phase was executed using Python in the FprojeNTI2.ipynb notebook.

**5.1. Data Exploration**

The dataset contains **1099 rows and 12 columns**.

* **Features:** Age, Gender, Stress Level, Academic Performance, Health Condition, Relationship Condition, Family Problem, Depression Level, Anxiety Level, Mental Support, Self Harm Story.
* **Target Variable:** Suicide Attempt.

**5.2. Data Cleaning**

1. **Missing Values:** The Family Problem column had 406 missing values (~40%). To preserve the feature, missing values were imputed with the string "Unknown".
2. **Duplicate Values:** Duplicate rows were identified but kept, as they could represent different individuals with similar conditions.
3. **Outliers:** An outlier in the Age column (age > 200) was detected and removed using the Interquartile Range (IQR) method.

**5.3. Data Preparation for Modeling**

1. **Categorical Feature Encoding:** OneHotEncoder was used to convert categorical features into a numerical format.
2. **Target Variable Transformation:** The Suicide Attempt target variable was mapped to numerical values: Never Thought -> 0, Thought -> 1, Attempted -> 2.

**6. Model Building and Evaluation**

**6.1. Handling Imbalanced Data**

The target variable's classes were imbalanced. To address this, **RandomOverSampler** and **RandomUnderSampler** from the imblearn library were used to create a balanced dataset for more effective model training.

**6.2. Models Implemented & Results**

The data was split into 80% for training and 20% for testing.

1. **Decision Tree:** Achieved an accuracy of **83%** on the imbalanced data. Performance improved significantly after data balancing and hyperparameter tuning (GridSearchCV).
2. **Random Forest:** This model proved to be a top performer. After balancing and tuning, it achieved an excellent accuracy of **88.2%**, with strong and balanced performance across all classes.
3. **Neural Network (NN):** A sequential neural network implemented with TensorFlow/Keras performed exceptionally well. After incorporating techniques like Dropout and BatchNormalization, it achieved a validation accuracy of approximately **91%**, making it a highly competitive model for this prediction task.

**7. Power BI Visualization and Analysis**

The BI track focused on translating raw data into actionable insights through an interactive dashboard designed to answer key stakeholder questions.

**7.1. Dashboard Content and Key Analytical Questions**

The final dashboard included the following pages, each addressing specific questions:

**A. Overview Page**  
This page provides a high-level summary of the dataset to answer foundational questions:

1. Do all depressed people attempt suicide?
2. How available is psychological support for people at risk of suicide?
3. Does academic performance vary among people at risk of suicide?
4. What is the social distribution (relationship status) of people at risk of suicide?
5. How many individuals are there by gender?

**B. Suicide Behavior Analysis Page**  
This page offers a deeper dive into the factors directly related to suicidal behavior:

1. Which gender attempts suicide the most?
2. Is there a relationship between suicide attempts and family problems?
3. What is the percentage of each suicide-related outcome (Never Thought, Thought, Attempted)?
4. Which gender receives the most support?
5. How do suicide rates vary by age and gender?
6. Does relationship status affect the suicide rate?

**C. Mental Health Distribution Page**  
This page analyzes the underlying mental health factors in the dataset:

1. What is the distribution of depression scores?
2. What is the distribution of anxiety scores?
3. What is the percentage of individuals who have a self-harm story?
4. What is the distribution of stress levels?
5. Which factors contribute the most to suicide attempts?
6. Which age group is more likely to self-harm under psychological pressure?

**8. Conclusion and Recommendations**

* **Conclusion:**  
  The project successfully analyzed a dataset on suicide attempts and developed effective predictive models. The **Random Forest** and **Neural Network** models delivered the highest accuracy and most balanced performance. The accompanying Power BI dashboard provides a powerful tool for stakeholders to explore the data and understand the key factors associated with suicide risk.
* **Future Recommendations:**
  1. **Enrich Data:** Acquire a larger, more diverse dataset to enhance model generalizability.
  2. **Advanced Models:** Experiment with algorithms like XGBoost or LightGBM for potential performance improvements.
  3. **Model Deployment:** Deploy the best-performing model as a web application or API for practical use in early-detection systems.
  4. **Text Analysis (NLP):** If text-based data becomes available, Natural Language Processing could be used to extract deeper qualitative insights into individuals' stories.