**Mental Health Prevalence Prediction Capstone Project**

**Title Page**

* Project: Mental Health Prevalence Capstone Project
* Author: Mary Gichara
* Date: September 2025

**Abstract**  
This project analyzes global mental health data to predict high depression prevalence. Using a Random Forest model, we achieved 85% accuracy and identified key demographic risk factors. The insights can help inform mental health interventions and policy decisions.

**Introduction**  
Depression affects millions worldwide, with varying prevalence across countries, age groups, and genders. This project aims to predict high-risk populations using historical prevalence data, providing actionable insights for mental health awareness and early intervention.

**Data**

* Dataset: 1-mental-illnesses-prevalence.csv
* Number of Rows: [6420]
* Number of Columns: [8]
* Key Features:
  + Country
  + Age Group
  + Sex
  + Depression prevalence
  + Other mental health metrics

Preprocessing steps included handling missing values, standardizing column names, and encoding categorical variables.

**Methodology**

1. **Exploratory Data Analysis (EDA)**
   * Identified distributions, trends, and correlations among features.
   * Visualized prevalence by country, age, and sex.
2. **Feature Engineering**
   * Created new variables if necessary.
   * Selected relevant features for modeling.
3. **Model Selection and Training**
   * Random Forest Classifier was chosen due to robustness.
   * Split data into training and testing sets.
4. **Model Evaluation**
   * Performance metrics calculated: accuracy, precision, recall, F1-score.
   * Visualized results using confusion matrix and feature importance plots.

**Results**

* Accuracy: 0.85
* Precision: 0.86
* Recall: 0.84
* F1 Score: 0.85

Visualizations:

* Include plots such as EDA charts, model evaluation metrics, and feature importance.

**Discussion / Interpretation**

* The Random Forest model effectively predicted high depression prevalence.
* Key predictors included [Age group, country and gender].
* Limitations: Dataset size, missing values, and possible biases.

**Conclusion**

* The project demonstrates that machine learning can help identify high-risk populations.
* Recommendations: Incorporate additional features, test other models, and use the insights to inform mental health interventions.

**Future Work**

* Compare multiple machine learning models (XGBoost, Logistic Regression)
* Incorporate external datasets for broader coverage
* Explore temporal trends and predictive analytics

**References**

* [Dataset source]
* [Articles or papers referenced]
* Python libraries used: pandas, NumPy, scikit-learn, matplotlib, seaborn
* Youtube