

Predicting Mental Health: Data Analysis and Model Development

Exploring mental illness prediction using synthetic depression data.

 by **TEMIJORUN GBOLAHAN**



Project Overview

Dataset

Synthetic Depression Dataset from Kaggle.

Approach

Data exploration, feature engineering, predictive modeling.

Tools

Python, SMOTE, Flask.



Data Preparation

Cleaning

Addressed missing values and duplicates.

Feature Selection

Significant associations with depression included.

Preprocessing

Used encoding and SMOTE for class imbalance.

MOLTAL AMI FE FAPLICE

MEKEAL POSICIH-REMHING GUIRRS INLSIENTS



Key Insights from Data Analysis

1

Lifestyle Impact

Parents show lower depression rates.

2

Unexpected Patterns

Non-smokers show higher depression rates.

3

Health and Social Factors

Employed individuals have higher depression rates.

Model Selection and Justification



- 1 Random Forest**
Chosen for interpretability and robustness.
- 2 XGBoost**
Tested but showed lower performance.
- 3 Enhancements**
Cross-Validation and SMOTE for accuracy.

MACHINE LEARNING

THE SCIENTIFIC MODEL CHOICES

Learn how to choose the right model for your data and problem. This guide covers the key factors to consider when selecting a machine learning model, from data characteristics to model performance.

ACCURACY
The ability of a model to correctly predict the outcome of a given input.

INTERPRETABILITY
The ability to understand and explain the model's predictions.

ROBUSTNESS
The ability of a model to maintain performance on new, unseen data.

SCALABILITY
The ability of a model to handle large volumes of data.

SPEED
The time taken by a model to train and make predictions.

MEMORY USAGE
The amount of memory required to store and run the model.

EASE OF USE
The simplicity of the model's interface and documentation.

FLEXIBILITY
The ability of a model to adapt to different data distributions and tasks.

COST
The financial and computational resources required to train and deploy the model.

SECURITY
The ability of a model to protect sensitive data and prevent unauthorized access.

PRIVACY
The ability of a model to ensure that user data is kept confidential and secure.

RELIABILITY
The consistency of a model's performance over time and across different environments.

MAINTAINABILITY
The ease of updating and maintaining the model as new data is added.

PORTABILITY
The ability of a model to run on different hardware and software platforms.

INTEROPERABILITY
The ability of a model to work seamlessly with other systems and data sources.

COMPATIBILITY
The ability of a model to work with different data formats and standards.

ACCESSIBILITY
The ability of a model to be used by a wide range of users and organizations.

SUSTAINABILITY
The ability of a model to be used for a long period of time without becoming obsolete.

ETHICALITY
The ability of a model to be used in a way that is fair, transparent, and accountable.

TRANSPARENCY
The ability of a model to provide clear and understandable explanations of its decisions.

ACCOUNTABILITY
The ability of a model to be held responsible for its actions and decisions.

TRUSTWORTHINESS
The ability of a model to be trusted by users and stakeholders.

CREDIBILITY
The ability of a model to be seen as a reliable source of information.

REPUTATION
The ability of a model to be seen as a positive and trustworthy entity.

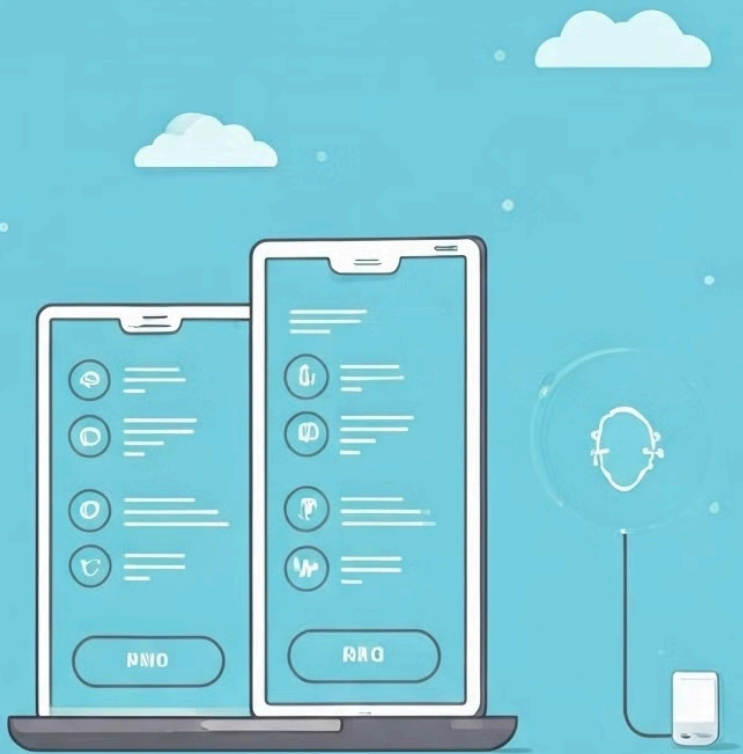
INFLUENCE
The ability of a model to have a positive impact on the world.

LEGACY
The ability of a model to leave a lasting and positive impact on future generations.



Model Performance Metrics

| Metric | Random Forest |
|-----------|---------------|
| Accuracy | ~62% |
| Precision | 59% |
| Recall | 62% |
| F1-Score | 60% |



Model Deployment and Endpoint



Flask API

Inputs demographic, lifestyle data.



JSON Structure

Field inputs like age, status.



Random Forest

Chosen for superior performance.



Model Limitations

Synthetic Data

Limited real-world generalizability.

Data Imbalance

May underperform in minority class.

Feature Engineering

Lacks genetic, psychological data.

Model Bias and Improvement Suggestions

Potential Biases

Socioeconomic and selection bias evident.

Improvement Strategies

Include psychological, environmental data.

Explainable AI

Use SHAP or LIME for transparency.



Conclusion

1

Summary

ML potential in mental health, focuses on depression.

2

Future Work

Refine with real-world data and advanced models.

3

Acknowledgments

Thanks to AXA Health and collaborators.