Predicting Mental Health: Data Analysis and Model Development

Exploring mental illness prediction using synthetic depression data.

by TEMILORUN 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.

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

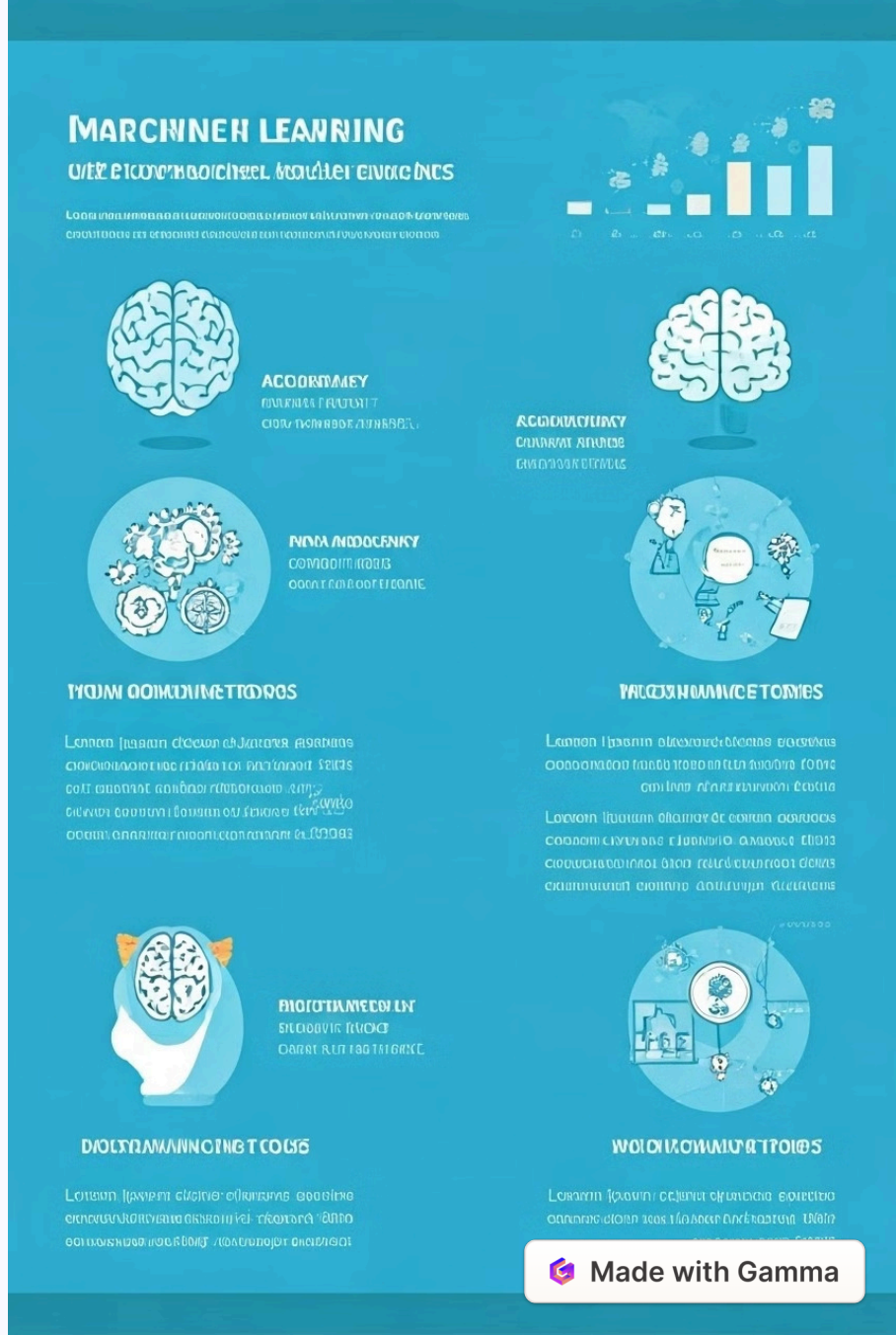
XGBoost

Tested but showed lower performance.

3

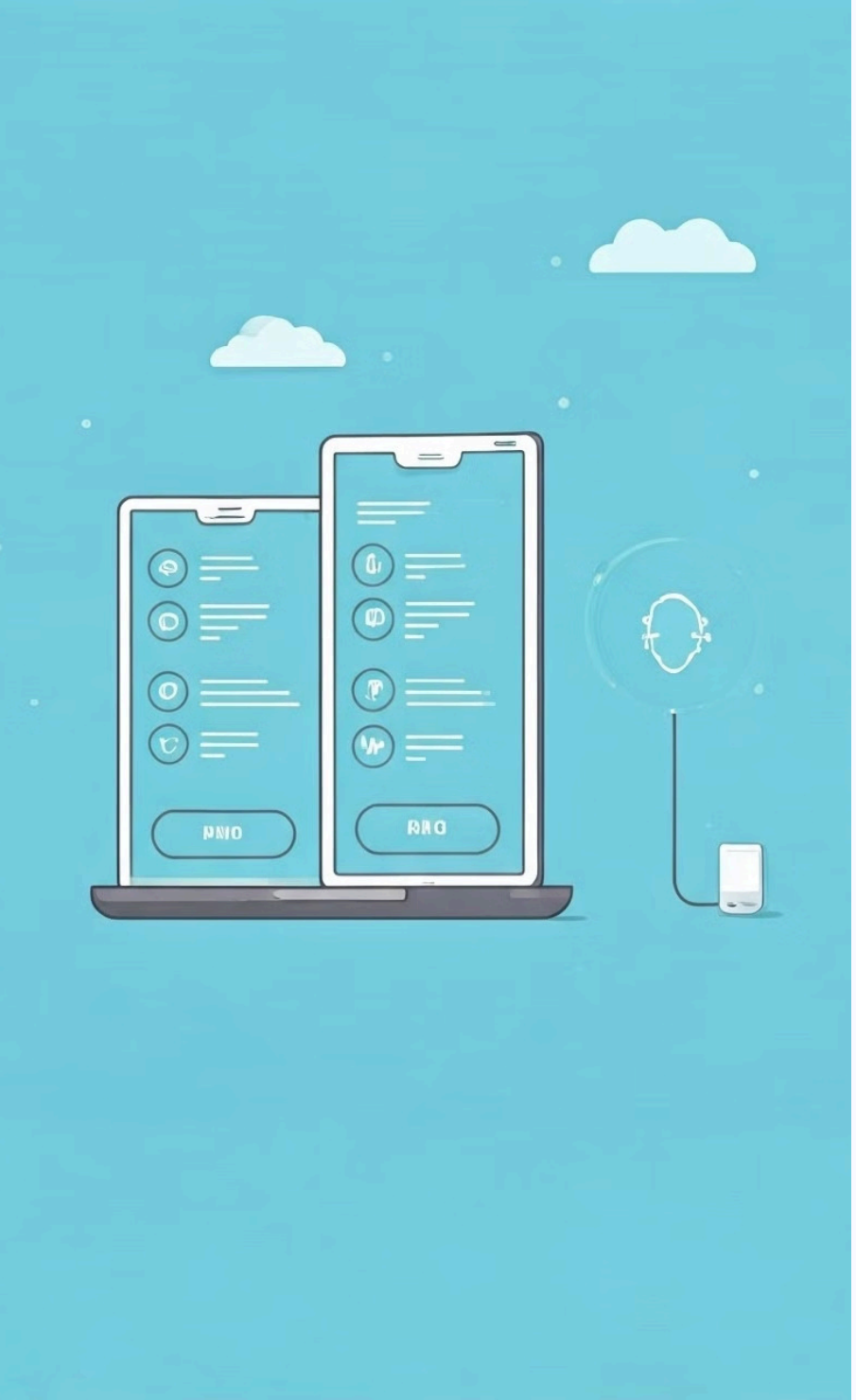
Enhancements

Cross-Validation and SMOTE for accuracy.



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.