Predicting Mental Health Using Machine Learning

Rohit Venugopal

Predicting Mental Health Issues Using Machine Learning

- This project was completed as part of a take-home task for a Data Scientist role at AXA Health.
- **The Goal:** To predict whether an individual will suffer from mental illness based on a dataset of demographic and behavioral factors.
- **Objective:** To deliver insights using machine learning to identify at-risk individuals and provide early interventions.

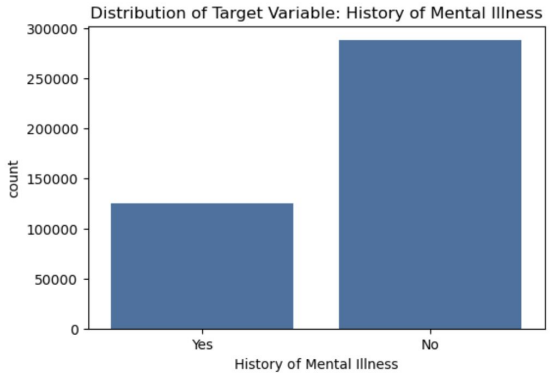
Dataset Overview

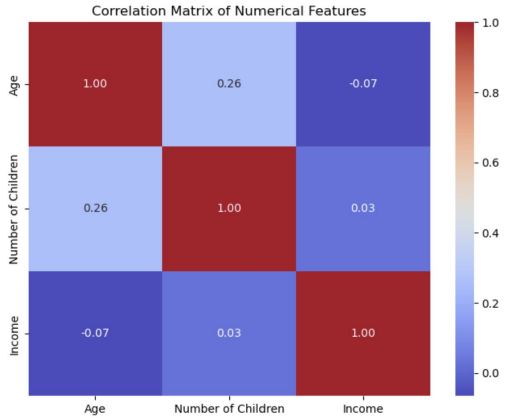
- **Source:** The dataset was downloaded from Kaggle (Depression Dataset).
- **Features:** 14 variables including demographics, lifestyle, and history.
- **Target:** “History of Mental Illness” (Yes/No).

Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	Sleep Patterns	History of Mental Illness	History of Substance Abuse	Family History of Depression	Chronic Medical Conditions
31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	26265.67	Moderate	Moderate	Fair	Yes	No	Yes	Yes
55	Married	High School	1	Non-smoker	Sedentary	Employed	42710.36	High	Unhealthy	Fair	Yes	No	No	Yes
78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	125332.79	Low	Unhealthy	Good	No	No	Yes	No
58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	9992.78	Moderate	Moderate	Poor	No	No	No	No
18	Single	High School	0	Non-smoker	Sedentary	Unemployed	8595.08	Low	Moderate	Fair	Yes	No	Yes	Yes

Exploratory Data Analysis

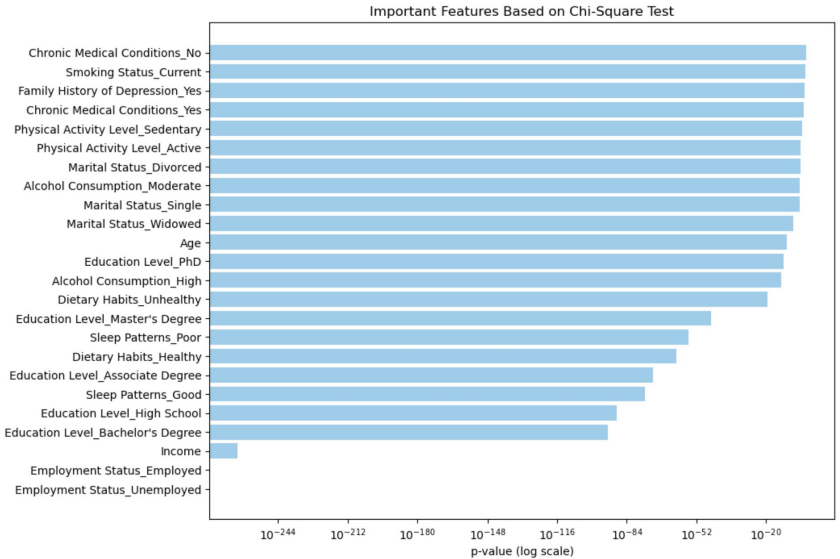
- Descriptive statistics were calculated for numeric and categorical features.
- Data was imbalanced, with more people reporting no history of mental illness than those who reported yes.
- Feature correlations were investigated using chi-square tests.





Feature Selection

- We included all variables except for “Name” (irrelevant for predictions).
- Features like marital status, income, employment status, and family history of depression had strong relationships with mental health outcomes.
- Categorical features were one-hot encoded, and numeric features were scaled.

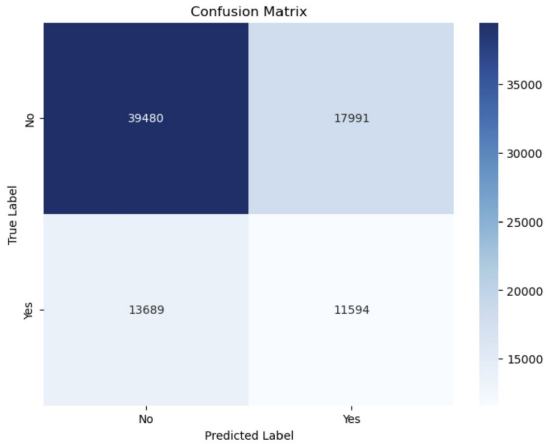


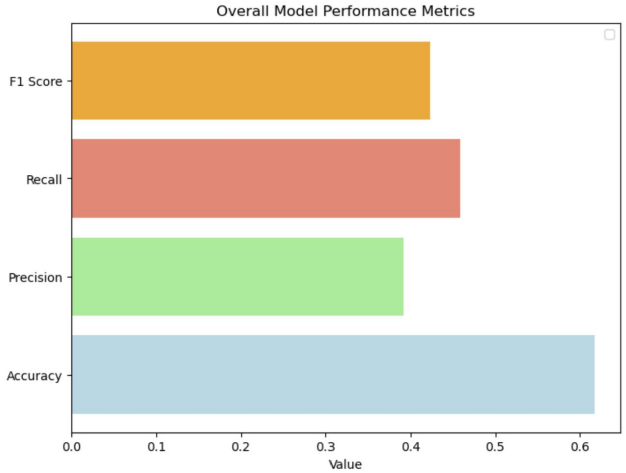
Model Development

- We explored several models: Decision Trees, Logistic Regression, Random Forest, and LightGBM.
- Cross-validation and hyperparameter tuning were used to select the best models.
- **SMOTE** was used to address the class imbalance in the training data
- Logistic Regression was chosen as the final model due to strong interpretability and balanced performance.

Model Performance

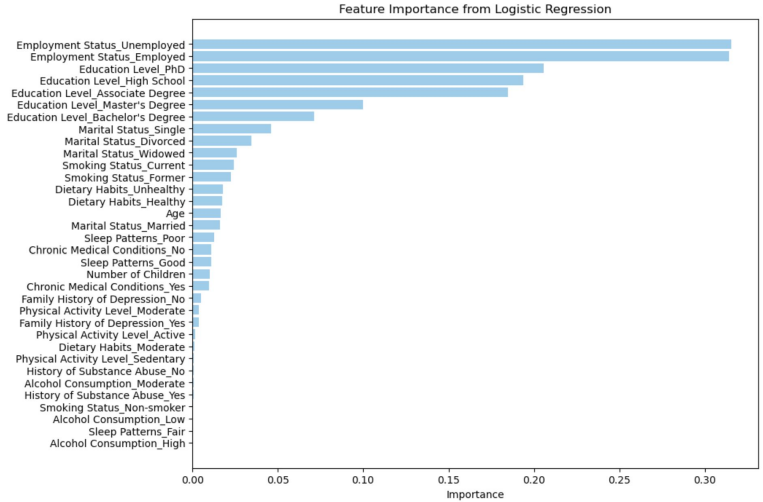
- The Logistic Regression model provided the best balance of precision (39%) and recall (46%).
- Accuracy: 61.7%, F1 Score: 42.3%.
- This model effectively identified individuals with mental illness, though improvements can be made in recall.





Key Insights

- Marital status, income, physical activity, and family history of depression were the strongest predictors.
- Addressing lifestyle factors and providing early interventions to individuals identified by the model can improve mental health outcomes.



Limitations & Improvements

- **Limitations**
  - The dataset has a significant class imbalance.
  - The model's recall could be further improved with additional features or better sampling techniques.
- **Improvements**
  - Collecting more balanced data.
  - Using advanced techniques like ensemble models or deep learning models for better accuracy and recall.
  - Potential inclusion of more behavioral and lifestyle variables.

Conclusion

- Logistic Regression provided an interpretable and reasonably accurate model for predicting mental health risks.
- This model can help healthcare professionals target early interventions for at-risk individuals, especially when combined with additional contextual data.
- The model is flexible and can be retrained as more data becomes available.