OCD Severity Prediction Using Machine Learning

Abstract

Obsessive-Compulsive Disorder (OCD) is a chronic and disabling psychiatric condition characterized by recurring intrusive thoughts (obsessions) and ritualistic behaviors (compulsions). Understanding and predicting the severity of OCD is vital for patient stratification, early intervention, and the development of personalized treatment strategies. In this project, a data-driven approach was used to analyze demographic and clinical features of OCD patients. Various statistical, visualization, and machine learning techniques were employed to identify patterns and predict the Total Y-BOCS Score, a standard clinical metric for OCD severity.

1. Introduction

The rising incidence and complexity of psychiatric disorders like OCD demand innovative approaches to diagnosis and treatment planning. Traditional methods often fall short in delivering precise prognostic evaluations. With the increasing availability of healthcare data, machine learning offers a promising alternative for deriving actionable insights and predictive outcomes.

The aim of this project was to explore the potential of machine learning in predicting OCD severity using a dataset of 1,500 patient records. This report covers the full analytical journey—from data cleaning and feature engineering to exploratory analysis, model training, evaluation, and clinical interpretation of results.

2. Dataset Overview

The dataset consisted of 1,500 patients and 34 variables, including: - Demographics: Age, Gender, Ethnicity, Marital Status, Education Level - Clinical: OCD Diagnosis Date, Duration of Symptoms, Obsession & Compulsion Types, Y-BOCS Scores - Comorbidities: Depression and Anxiety Diagnoses - Treatments: Medications (SSRIs, SNRIs), Previous Diagnoses (MDD, PTSD, Panic Disorder)

After preprocessing, the dataset had no critical missing values and all variables were appropriately encoded.

3. Data Preprocessing

3.1 Column Cleaning and Formatting

- All column names were standardized by removing spaces and special characters.

3.2 Missing Values Handling

- ‘Medications’ and ‘Previous Diagnoses’ were filled with ‘None’
- Records missing 'Age', 'Gender', or 'Total Y-BOCS Score' were dropped

3.3 Encoding

- Binary categorical variables were label-encoded
- Multiclass categorical variables were one-hot encoded

3.4 Feature Engineering

- A new column, `Total_YBOCS_Score`, was created as the sum of obsession and compulsion scores.

4. Exploratory Data Analysis (EDA)

4.1 Statistical Summary

- **Median Age**: 47 years
- **Median OCD Severity Score**: 40
- **Standard Deviation of Total Score**: ~16.95

4.2 Distribution Visualizations

- **Age**: Bell-shaped distribution centered around 45–50 years
- **Severity**: Most patients scored between 30 and 45
- **Gender Distribution**: Balanced male and female counts
- **Ethnicity, Marital Status, and Education**: Diverse and reasonably distributed

4.3 Boxplots and Correlations

- Severity tended to be higher in patients with depression, anxiety, and a family history of OCD
- Highest correlation was found between obsession and compulsion scores and the Total Y-BOCS Score ($r \approx 0.72$)

5. Model Building and Evaluation

5.1 Models Used

- **Random Forest Regressor**: Non-linear ensemble model
- **Linear Regression**: Baseline for comparison

5.2 Performance Metrics (Random Forest)

- **MAE**: 0.57
- **MSE**: 0.60
- **RMSE**: 0.77
- **R^2 (Test Set)**: 0.9979
- **Cross-Validated R^2 Mean**: 0.9978

5.3 Feature Importance (Top 5)

1. **Y-BOCS Score (Obsessions)** – 50.2%
2. **Y-BOCS Score (Compulsions)** – 49.4%

3. **Duration of Symptoms** – 0.07%
4. **Age** – 0.06%
5. **Anxiety Diagnosis** – 0.01%

5.4 Interpretation

The severity score is overwhelmingly driven by the component Y-BOCS scores, validating their clinical role. Demographic and comorbidity factors had minimal influence.

6. Clinical and Demographic Insights

6.1 Medication Influence

• SSRIs and SNRIs were frequently prescribed and associated with higher severity

6.2 Depression & Anxiety

• Patients diagnosed with comorbid depression or anxiety consistently showed higher severity scores

6.3 Age and Gender Trends

• Female patients showed a wider range in scores
• Older patients had more variability in obsessive patterns

6.4 Obsession and Compulsion Types

• Harm-related and symmetry obsessions were prominent
• Washing and counting were the most frequent compulsions

7. Deployment and Utility

The Random Forest model was saved as a `.pkl` file using joblib, enabling real-time predictions in applications or web interfaces. This opens the door to clinical decision support tools that could assess severity instantly based on patient input.

8. Limitations

• The dataset, though extensive, was confined to a single demographic region
• Some categorical classes (like rare medications) were underrepresented
• Data relied heavily on self-reporting, which can introduce bias

9. Conclusion

This study demonstrates how machine learning can effectively predict OCD severity using structured demographic and clinical data. The project followed a rigorous data science process: - Cleaned and transformed raw data - Extracted clinical insights via EDA - Trained and evaluated powerful ML models - Interpreted features in clinical context

Final Remarks

The model's near-perfect R^2 scores suggest high predictability, but real-world applications should be approached cautiously. Further research should include more dynamic features (e.g., treatment duration), psychological assessments, and multi-institutional data for generalization.

This project not only provides a template for psychiatric predictive modeling but also reinforces the synergy between machine learning and mental health research.

Keywords: OCD, Machine Learning, Y-BOCS, Data Science, Mental Health, Random Forest, Clinical Prediction, EDA, Feature Importance