OCD Patient Dataset: Demographics & Clinical Data

Presented by: Yash Sahu

Table of Content

1. Introduction & Problem Statement
2. Dataset Description
3. Techniques Used
4. Model Training & Evaluation
5. Output & Results

Introduction & Problem Statement

Introduction

- OCD is a chronic disorder characterized by persistent obsessions and compulsions.
- The dataset of 1,500 patients includes demographics, symptom duration, Y-BOCS scores, comorbidities, and treatments to analyze OCD patterns.

Problem Statement

- Identifying risk factors and patterns in OCD.
- Understanding the impact of comorbidities like depression and anxiety.
- Evaluating treatment effectiveness based on medications.

Dataset Description

Overview

- The dataset contains 1,500 OCD patients, covering demographics, symptom duration, Y-BOCS scores, comorbidities, and medications.
- Key attributes include age, gender, ethnicity, marital status, obsession/compulsion types, and treatment history.

Data Insights

- Total Records: 1,500 rows and 17 columns.
- There are 248 null values in Previous Diagnoses feature and 386 null values in Medications feature.

Dataset Description

```
df.head()
```
✓ 0.0s Python

	Patient ID	Age	Gender	Ethnicity	Marital Status	Education Level	OCD Diagnosis Date	Duration of Symptoms (months)	Previous Diagnoses	Family History of OCD	Obsession Type	Compulsion Type	Y-BOCS Score (Obsessions)	Y-BOCS Score (Compulsions)	Depressi Diagno
0	1018	32	Female	African	Single	Some College	2016-07-15	203	MDD	No	Harm-related	Checking	17	10	
1	2406	69	Male	African	Divorced	Some College	2017-04-28	180	NaN	Yes	Harm-related	Washing	21	25	
2	1188	57	Male	Hispanic	Divorced	College Degree	2018-02-02	173	MDD	No	Contamination	Checking	3	4	
3	6200	27	Female	Hispanic	Married	College Degree	2014-08-25	126	PTSD	Yes	Symmetry	Washing	14	28	
						High	2022-02-								

This is the output of df.head() first five rows of dataset.

```
### list of columns
df.columns
```
✓ 0.0s

```
Index(['Patient ID', 'Age', 'Gender', 'Ethnicity', 'Marital Status',
       'Education Level', 'OCD Diagnosis Date',
       'Duration of Symptoms (months)', 'Previous Diagnoses',
       'Family History of OCD', 'Obsession Type', 'Compulsion Type',
       'Y-BOCS Score (Obsessions)', 'Y-BOCS Score (Compulsions)',
       'Depression Diagnosis', 'Anxiety Diagnosis', 'Medications'],
      dtype='object')
```

Techniques Used

Introduction

- **Exploratory Data Analysis (EDA):** Identified trends in OCD severity, comorbidities, and treatment effectiveness.
- **Machine Learning Models:** Tested classification models for predicting OCD severity and treatment outcomes.

Best Model Performance:

- After testing, AdaBoostClassifier provided the highest accuracy:
- **Accuracy Score: 0.54**

Techniques Used

Model Evaluation Metrics

- **Data Preprocessing:** Handled missing values, performed encoding, and feature scaling.
- **Model Performance:** Evaluated accuracy using metrics like Accuracy Score, Confusion Matrix, F1-score, Precision and Recall scores for classification tasks.

Model Training & Evaluation

Model Training Process:

1. **Data Splitting:**

- Dataset was split into 80% training data and 20% test data

2. **Feature Engineering & Preprocessing:**

- One-Hot Encoding applied to categorical variables (e.g., Gender, Ethnicity, Obsession Type).
- Feature Scaling performed where necessary (e.g., Y-BOCS Scores, Duration of Symptoms).

3. **Model Training:**

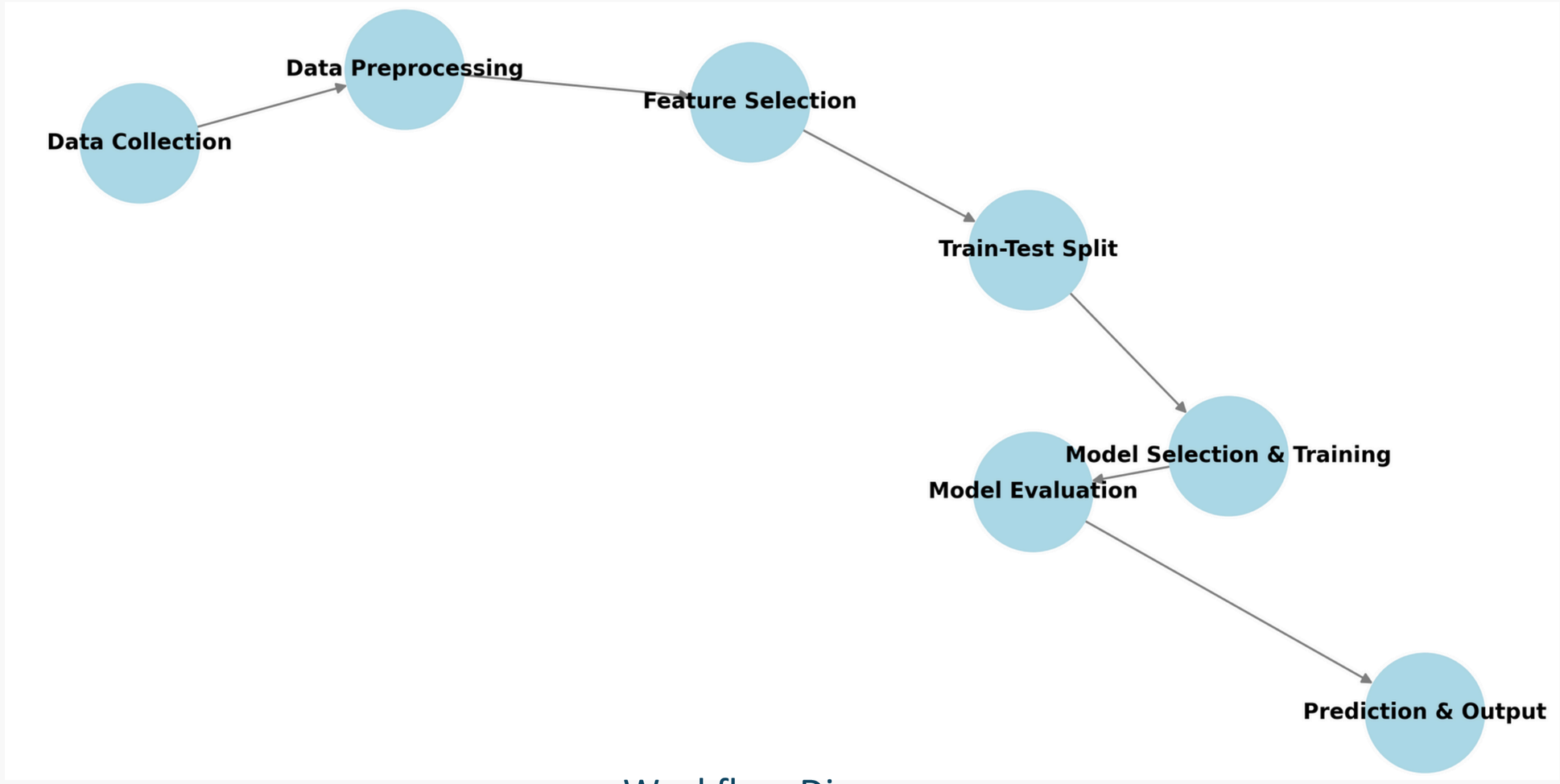
- Tested multiple algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost etc).
- Ada Boost Classifier performed the best in predicting OCD severity and treatment outcomes.

Model Training & Evaluation

Workflow Steps

1. Data Collection – Gather OCD patient demographics and clinical data.
2. Data Preprocessing – Handle missing values, apply encoding for categorical variables, and scale numerical features.
3. Feature Selection – Identify key features affecting OCD severity (e.g., Y-BOCS scores, comorbidities, medication use).
4. Train-Test Split – Split dataset into 80% training and 20% testing for model evaluation.
5. Model Selection & Training – Train multiple models (Logistic Regression, Decision Tree, Random Forest, XGBoost etc).
6. Model Evaluation – Compare accuracy, F1-score, and confusion matrix to select the best model.
7. Prediction & Output – Use the trained model to predict OCD severity and treatment effectiveness.

Model Training & Evaluation



Workflow Diagram

Output & Results

Web Application

- The **Ada Boost Classifier** was the best-performing model.
- **Accuracy:**
- Before Hyperparameter Tuning: 0.51
- After Hyperparameter Tuning: 0.54

Observations

- Model predicts with high accuracy.
- Some slight variations due to feature importance & dataset limitations.
- Further tuning could improve generalization.

Thank You!