DATA SCIENCE INTERNSHIP

Project 2 - OCD Patient Dataset

UMID03072548194

Anjali Shibu

anjalishi1994@gmail.com

Contents

Introduction

The given dataset contains demographic and clinical data of OCD patients. The file is downloaded as a csv file. Python programming is used in Colab environment for this project. We conduct a detailed analysis of the data to gain insights from it.

Import necessary python libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import pandas as pd
import matplotlib.pyplot as plt
import shap
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error
from sklearn.cluster import KMeans
import joblib
```

```
[ ] df = pd.read_csv('/content/OCD Patient Dataset_ Demographics & Clinical Data.csv')
```

We imported the libraries and dataset is loaded as pandas dataframe.

Variable description

```
df.head()
```

	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
0	1018	32	Female	African	Single	Some College	15-07-2016	203	MDD	No	Harm-related	Checking	17	10	Yes	Yes	SNRI
1	2406	69	Male	African	Divorced	Some College	28-04-2017	180	NaN	Yes	Harm-related	Washing	21	25	Yes	Yes	SSRI
2	1188	57	Male	Hispanic	Divorced	College Degree	02-02-2018	173	MDD	No	Contamination	Checking	3	4	No	No	Benzodiazepine
3	6200	27	Female	Hispanic	Married	College Degree	25-08-2014	126	PTSD	Yes	Symmetry	Washing	14	28	Yes	Yes	SSRI
4	5824	56	Female	Hispanic	Married	High School	20-02-2022	168	PTSD	Yes	Hoarding	Ordering	39	18	No	No	NaN

In the dataset there are 17 variables. Variable description is as follows:

Variable description

- Patient ID: Unique identifier for each patient.
- Age: Age of the patient.
- Gender: Gender of the patient.
- Ethnicity: Ethnicity of the patient.

- Marital Status: Marital status of the patient.
- Education Level: Level of education attained by the patient.
- OCD Diagnosis Date: Date when OCD was diagnosed.
- Duration of Symptoms (months): Duration for which the patient has been experiencing symptoms.
- Previous Diagnoses: Any previous diagnoses before OCD.
- Family History of OCD:Whether the patient has a family history of OCD.
- Obsession Type: Type of obsessions experienced by the patient.
- Compulsion Type: Type of compulsions experienced by the patient.
- Y-BOCS Score (Obsessions): Y-BOCS score related to obsessions.
- Y-BOCS Score (Compulsions): Y-BOCS score related to compulsions.
- Depression Diagnosis: Whether the patient has been diagnosed with depression
- Anxiety Diagnosis: Whether the patient has been diagnosed with anxiety.
- Medications: Medications the patient is currently taking

Check missing data

```
missing_data = df.isnull().sum()
print(missing_data)
```

```
Patient ID                         0
Age                                0
Gender                             0
Ethnicity                          0
Marital Status                     0
Education Level                    0
OCD Diagnosis Date                 0
Duration of Symptoms (months)      0
Previous Diagnoses               248
Family History of OCD              0
Obsession Type                     0
Compulsion Type                    0
Y-BOCS Score (Obsessions)          0
Y-BOCS Score (Compulsions)         0
Depression Diagnosis               0
Anxiety Diagnosis                  0
Medications                      386
dtype: int64
```

We find that some data is missing in the variable ‘Previous diagnoses’ (248 values)and ‘Medications’(386 values).

```
import pandas as pd

# df is DataFrame,create a copy named df_clean
df_clean = df.copy()

# Add 'Unknown' to the categories of 'Previous Diagnoses' and 'Medications'
for col in ['Previous Diagnoses', 'Medications']:
    if df_clean[col].dtype.name == 'category':
        # Add 'Unknown' to the categories
        df_clean[col] = df_clean[col].cat.add_categories(['Unknown'])

    # Impute missing values with 'Unknown'
    df_clean[col] = df_clean[col].fillna('Unknown')

# Verify no missing values remain
print(df_clean[['Previous Diagnoses', 'Medications']].isnull().sum())
```

```
Previous Diagnoses    0
Medications           0
dtype: int64
```

We then create a copy of the original dataframe(figure above), checks if the columns are categorical,adds a new category called unknown to both columns,fills missing values in those columns with 'unknown', verifies that there are no missing values left by printing the count of nulls.

This makes sure no data is lost due to missing values.The model can still use these columns without errors .

Repeating values

```
for column in df.columns:
    duplicates = df_clean[column][df_clean[column].duplicated()]
    if not duplicates.empty:
        print(f"Column '{column}' has repeated values.")
        print(duplicates.unique())
        print('-' * 50)
    else:
        print(f"Column '{column}' has all unique values.")
```

The above code shows repeating values . From the output we see a lot of repeating values in all the variables in the data. Here when we examine the repeating values ,we see Patient ID is repeating.

```
# Find and display duplicate 'Patient ID' values
duplicate_patient_ids = df_clean[df_clean['Patient ID'].duplicated(keep=False)]
display(duplicate_patient_ids)
```

The code gives an output of 206 rows repeating.

Data cleaning

```
patient_details = df_clean[df_clean['Patient ID'] == 1018]
display(patient_details)
```

	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
0	1018	32	Female	African	Single	Some College	15-07-2016	203	MDD	No	Harm-related	Checking	17	10	Yes	Yes	SNRI
209	1018	73	Female	Caucasian	Single	College Degree	05-12-2015	150	Panic Disorder	Yes	Symmetry	Counting	9	15	No	No	SNRI

When we check a particular Patient ID we see that though the ID is same, the data is entirely different,so deleting the duplicate Patient ID can result in losing critical patient data.Hence we decide to drop Patient ID variable.

```
df_clean = df_clean.drop(columns = ['Patient ID'])
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             1500 non-null   int64
 1   Gender                          1500 non-null   object
 2   Ethnicity                       1500 non-null   object
 3   Marital Status                  1500 non-null   object
 4   Education Level                 1500 non-null   object
 5   OCD Diagnosis Date              1500 non-null   object
 6   Duration of Symptoms (months)   1500 non-null   int64
 7   Previous Diagnoses              1500 non-null   object
 8   Family History of OCD           1500 non-null   object
 9   Obsession Type                  1500 non-null   object
 10  Compulsion Type                 1500 non-null   object
 11  Y-BOCS Score (Obsessions)       1500 non-null   int64
 12  Y-BOCS Score (Compulsions)      1500 non-null   int64
 13  Depression Diagnosis            1500 non-null   object
 14  Anxiety Diagnosis               1500 non-null   object
 15  Medications                     1500 non-null   object
dtypes: int64(4), object(12)
memory usage: 187.6+ KB
```

After dropping the variable Patient ID we have 16 variables left.

Change data type

```
object_cols = ['Gender', 'Ethnicity', 'Marital Status', 'Education Level','Previous Diagnoses',
               'Family History of OCD', 'Obsession Type', 'Compulsion Type',
               'Depression Diagnosis', 'Anxiety Diagnosis', 'Medications']

for col in object_cols:
    df_clean[col] = df_clean[col].astype('category')

df_clean['OCD Diagnosis Date'] = pd.to_datetime(df_clean['OCD Diagnosis Date'])

df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 16 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Age                           1500 non-null   int64
 1   Gender                        1500 non-null   category
 2   Ethnicity                     1500 non-null   category
 3   Marital Status                1500 non-null   category
 4   Education Level               1500 non-null   category
 5   OCD Diagnosis Date            1500 non-null   datetime64[ns]
 6   Duration of Symptoms (months)  1500 non-null   int64
 7   Previous Diagnoses            1500 non-null   category
 8   Family History of OCD         1500 non-null   category
 9   Obsession Type                1500 non-null   category
 10  Compulsion Type               1500 non-null   category
 11  Y-BOCS Score (Obsessions)     1500 non-null   int64
 12  Y-BOCS Score (Compulsions)    1500 non-null   int64
 13  Depression Diagnosis          1500 non-null   category
 14  Anxiety Diagnosis             1500 non-null   category
 15  Medications                   1500 non-null   category
dtypes: category(11), datetime64[ns](1), int64(4)
memory usage: 76.7 KB
/tmp/ipython-input-13-456675430.py:8: UserWarning: Parsing dates in %d-%m-%Y format when dayfirst=False (the default) was specified. Pass `dayfirst=True` or specify a format to silence this warning.
  df_clean['OCD Diagnosis Date'] = pd.to_datetime(df_clean['OCD Diagnosis Date'])
```

Some object type variables are converted into category type and diagnosis date is converted to date time type.

Unique values

```
selected_cols = [
    'Gender', 'Ethnicity', 'Marital Status', 'Education Level', 'Previous Diagnoses',
    'Family History of OCD', 'Obsession Type', 'Compulsion Type',
    'Depression Diagnosis', 'Anxiety Diagnosis', 'Medications'
]

for col in selected_cols:
    print(f"Unique values in '{col}':")
    print(df_clean[col].unique())
    print('-' * 150)
```

The above given code displays unique values in each variables.

We see there are 2 values in gender- male and female.

4 values in ethnicity- African , Hispanic, Asian, Caucasian.

3 values in marital status – Single,divorced,married.

4 values in education level – college degree,graduate degree,high school, some college

5 values in previous diagnoses – GAD,MDD,PTSD,Panic disorder , Unknown

2 values in family history of OCD- No, yes

5 values in obsession type – Contamination, harm-related, hoarding , religious , symmetry.

5 values in compulsion type – Checking, counting , ordering , praying ,washing.

2 unique values in depression diagnosis , anxiety diagnoses - no, yes.

4 unique values in medications – Benzodiazepine , SNRI , SSRI , unknown

Some of these variables are not known to people from other than medical background,so an explanation of these variables and attributes are given for clear understanding

Obsession type: It shows the kind of upsetting thoughts people with OCD often get stuck on,like fears or worries they can't shake,that lead them to do repetitive actions to feel better.

Contamination: Fear of germs, dirt, or illness

Harm-related: Unwanted thoughts of causing injury to oneself or others

Hoarding: Difficulty discarding items, fear of losing something important

Religious: Intrusive blasphemous thoughts or fears of moral wrongdoing

Symmetry: Intense need for order or exactness

Compulsion type : It shows the kind of actions people with OCD feel they have to do,like cleaning, checking, or repeating things,to calm their anxious thoughts. These behaviors usually follow fixed patterns and feel hard to control.

Washing: Excessive hand-washing or cleaning

Checking: Repeatedly verifying things (e.g. locks, appliances)

Counting: Mental or physical counting routines

Praying: Ritualistic religious recitations

Ordering: Arranging items in a specific manner or alignment

Y-BOCS Score (Obsessions)

The Yale-Brown Obsessive Compulsive Scale (Y-BOCS) is a standard clinical tool to assess the severity of OCD symptoms. This column records numeric scores evaluating:

Time consumed by obsessive thoughts

Level of distress and interference

Resistance and control over the obsessions

Higher values (up to 40) suggest more intense obsession symptoms.

Y-BOCS Score (Compulsions)

This score quantifies the severity of compulsive behaviors using the same scale and methodology as for obsessions. It accounts for:

Frequency and duration of compulsive acts

Degree of effort to resist

Level of control and impairment

Together, the obsession and compulsion scores provide a comprehensive metric of OCD severity.

Medications:

Benzodiazepine: A class of sedative drugs often used short-term to reduce intense anxiety. They act quickly but carry a risk of dependency, so doctors prescribe them cautiously.

SNRI (Serotonin-Norepinephrine Reuptake Inhibitor): Antidepressants that increase levels of serotonin and norepinephrine—two key mood-regulating chemicals. Helpful for treating depression and anxiety linked to OCD.

SSRI (Selective Serotonin Reuptake Inhibitor): The most commonly used medication for OCD. SSRIs boost serotonin levels in the brain and help reduce obsessions and compulsions over time.

Previous diagnostics:

GAD- Generalized Anxiety Disorder—chronic worry and nervousness about everyday life

MDD- Major Depressive Disorder—persistent sadness, hopelessness, and low energy

PTSD- Post-Traumatic Stress Disorder—distressing symptoms following a traumatic event

Panic Disorder- Sudden, intense episodes of fear and physical symptoms like heart palpitations

Feature engineering

```
df_clean['Total Y-BOCS Score'] = df_clean['Y-BOCS Score (Obsessions)'] + df_clean['Y-BOCS Score (Compulsions)']
df_clean['Age Group'] = pd.cut(df_clean['Age'], bins=[0, 30, 50, 100], labels=['Young', 'Middle', 'Older'])
df_clean.info()
```

Total Y-BOCS Score was created by summing the scores from obsessions and compulsions. This provides a unified metric to quantify the overall severity of OCD symptoms.

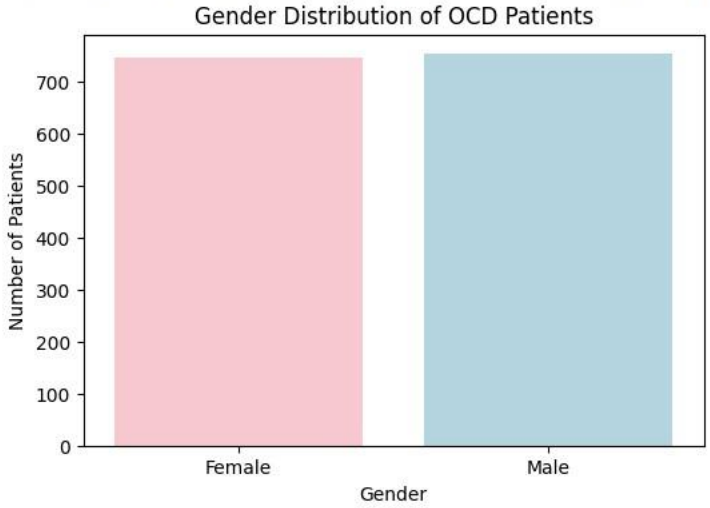
Age binning- Age group- Patients were categorized into three age groups using binning. This transformation helps in stratifying patients for demographic analysis and modelling , especially when age has non-linear effects.

These will be used in visualizations.

EDA

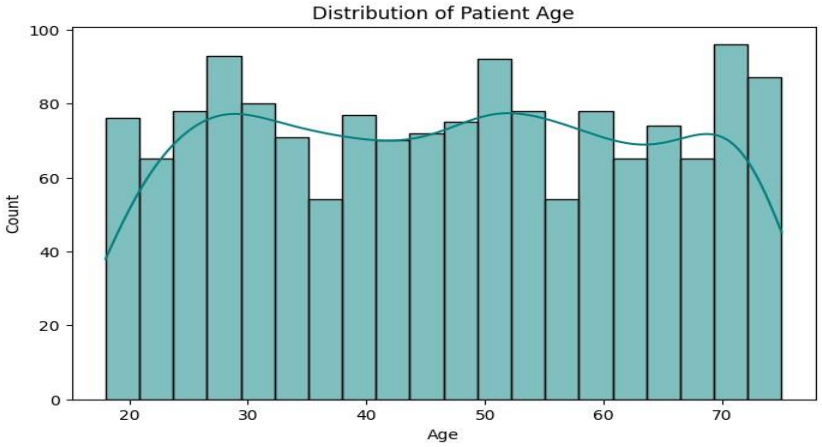
Male-female distribution

```
Gender
Male      753
Female    747
Name: count, dtype: int64
/tmp/ipython-input-16-3643158082.py:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in

  sns.barplot(x=gender_counts.index, y=gender_counts.values, palette=gender_pal
```



The dataset includes a balanced representation of genders among OCD Patients with 50.2% males and 49.8% females.This minimises potential gender bias in subsequent analysis.

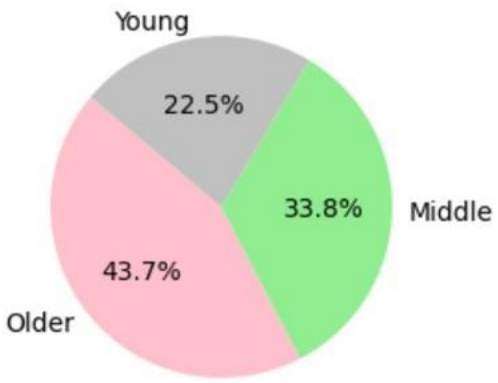
Patient age distribution



Bar chart was plotted showing the distribution of patient age . X-axis : Age range and Y-axis : count of patients in each age group. The chart includes a smooth trend line overlaying the bars to highlight peaks in age distribution.

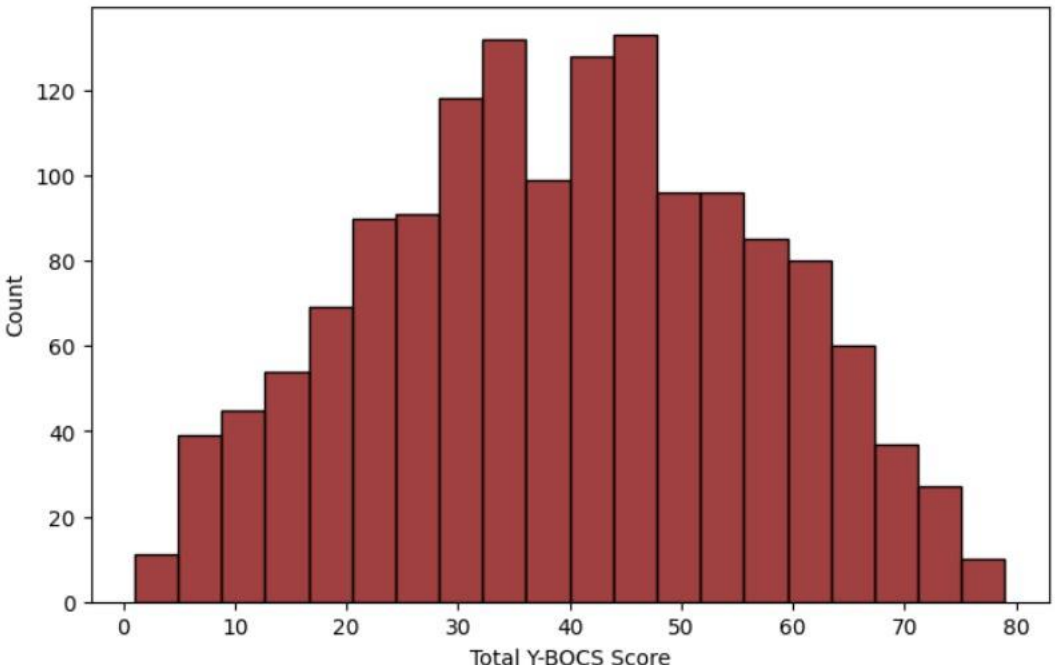
Patient counts are relatively distributed evenly across age groups. Notable peaks are observed around ages 30,50 and 70 suggesting higher prevalence or diagnosis rates in these age brackets.

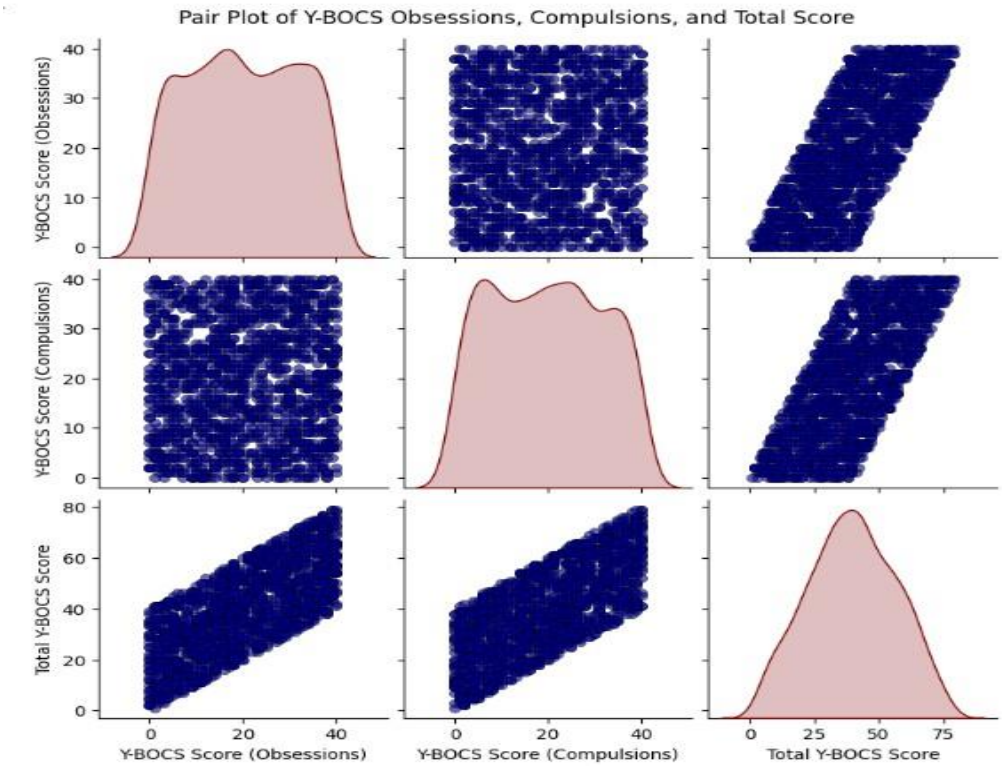
Pie chart – Distribution of patients by age group



The dataset reveals a diverse age distribution with ‘Middle’ aged individuals forming the largest group at 43.7%. The ‘Young’ and ‘ Older’ categories account for 22.5% and 33.8% of the data respectively. Notably , the ‘Older’ group is proportionally larger than the ‘ Young’ group , suggesting potential skewness or specific demographic trends in the dataset.

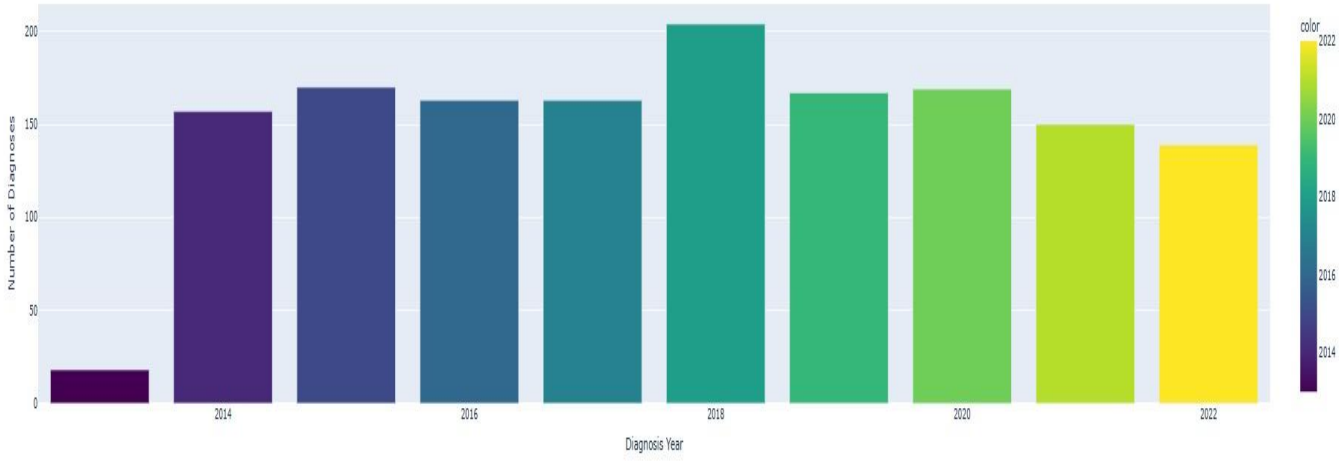
Distribution of total Y-BOCS score





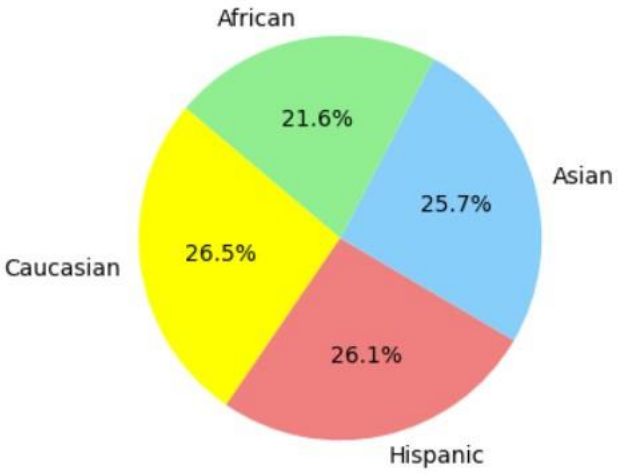
The pair plot reveals a strong linear relationship between Y-BOCS Obsessions and compulsions scores, validating the use of a Total Y-BOCS score as a combined measure of symptom severity. Both subscales exhibit right-skewed distributions , suggesting most patients in this cohort have mild-to-moderate symptoms. Notably, the total score's distribution shows a peak at moderate ranges (20-40), with fewer cases in the severe range (>50). Outliers like high obsessions with low compulsions may warrant further clinical review.

Frequency of OCD Diagnoses by year



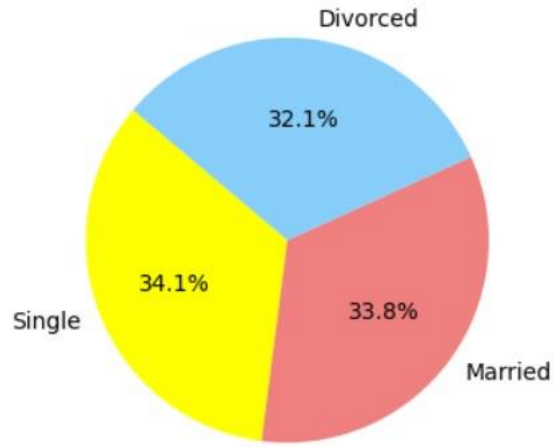
The diagnosis frequency shows a clear upward trend from 2013-2018, peaking at 204 cases in 2018 – a 13% increase from the previous year . Post- 2018 , volume stabilize near 160 cases /year , with a modest decline during pandemic years (2020-2022). The outlier low count in 2013 likely reflects incomplete historical records rather than true epidemiological variation.

Ethnicity distribution of OCD patients



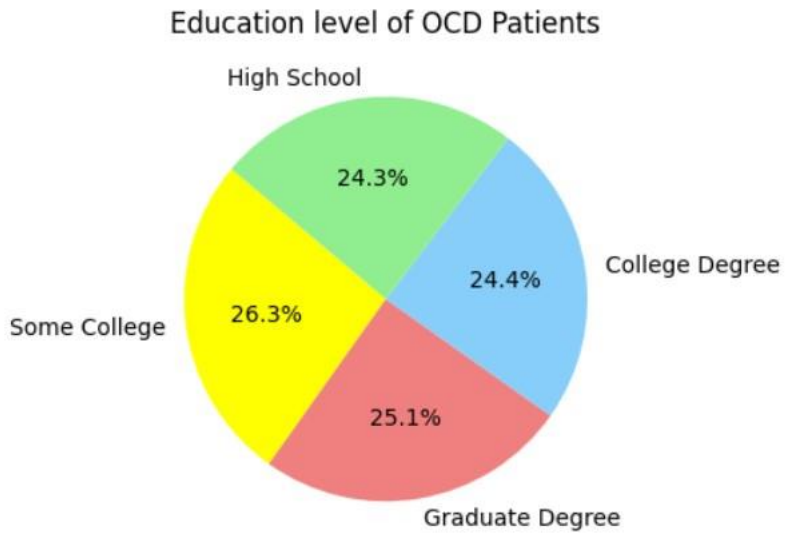
From the above pie chart , the dataset demonstrates relatively balanced ethnic representation , with all major groups comprising 21-27% cases . Caucasian(26.5%) and Hispanic (26.1%) individuals show marginally higher representation than African (26.1%) and Asian (25.7%) groups. This distribution suggests no strong ethnic sampling bias .

Marital status of OCD patients



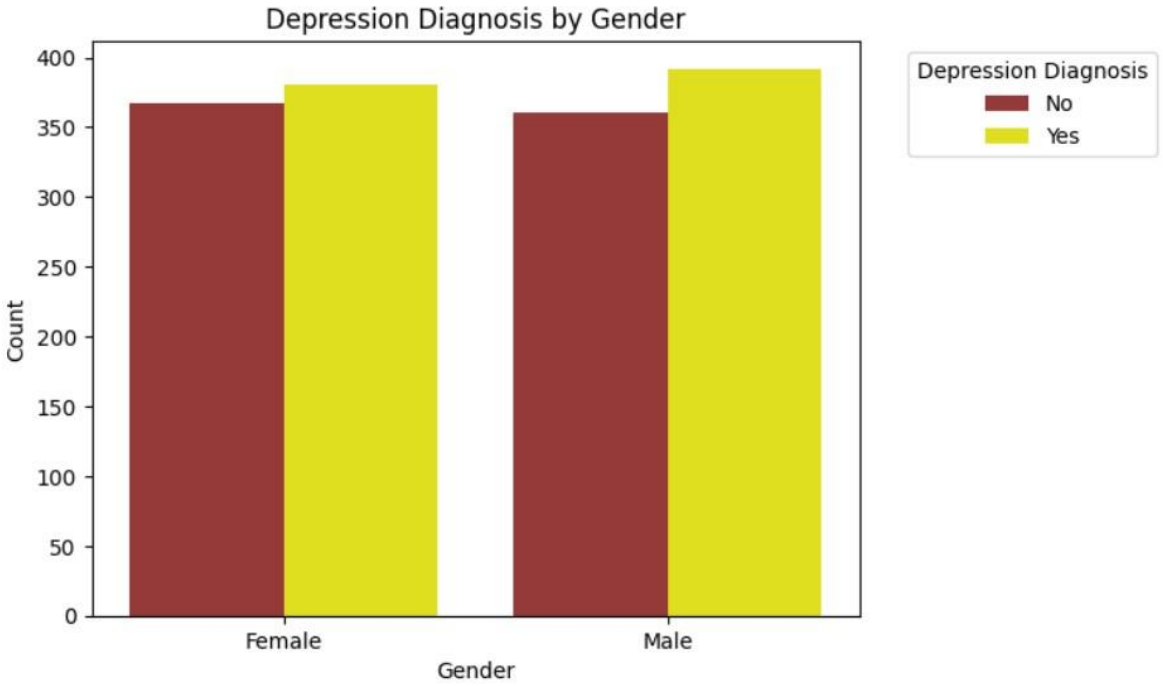
The sample shows nearly identical representation across marital statuses , with single (34.1%) , married (33.8%) and divorced (32.1%) individuals. This balanced distribution indicates that OCD diagnosis in this cohort is independent of marital status.

Education level of OCD Patients



The education distribution reveals that most OCD patients (75.7%) have at least some college experience , with largest groups being those with some college (26.3%) and college degrees (24.4%). Notably, 24.3% completed only high school, while 25.1% hold graduate degrees.OCD occurs across all education levels , highlighting the need for inclusive patient education materials.

Depression diagnosis by gender

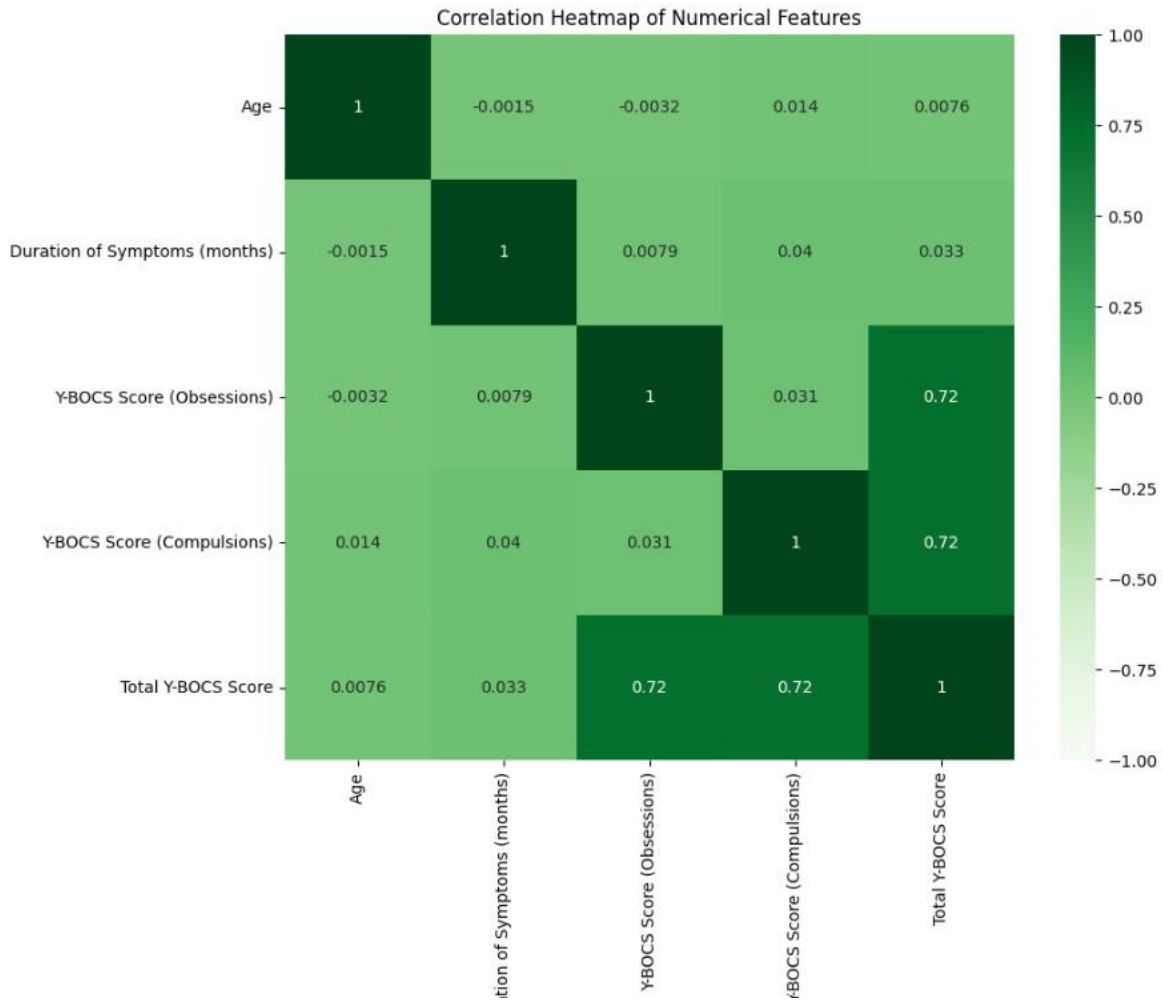



```
Chi-square = 0.21, p = 0.995
Expected frequencies (no association):
 [[ 362.544  384.456  747.   ]
 [ 365.456  387.544  753.   ]
 [ 728.     772.    1500.   ]]
```

A chi-square test of independence found no significant association between gender and depression diagnosis . The observed frequencies of depression were almost identical to expected frequencies

under the null hypothesis, indicating that depression rates did not differ meaningfully between male and female OCD patients in this cohort.

Correlation analysis

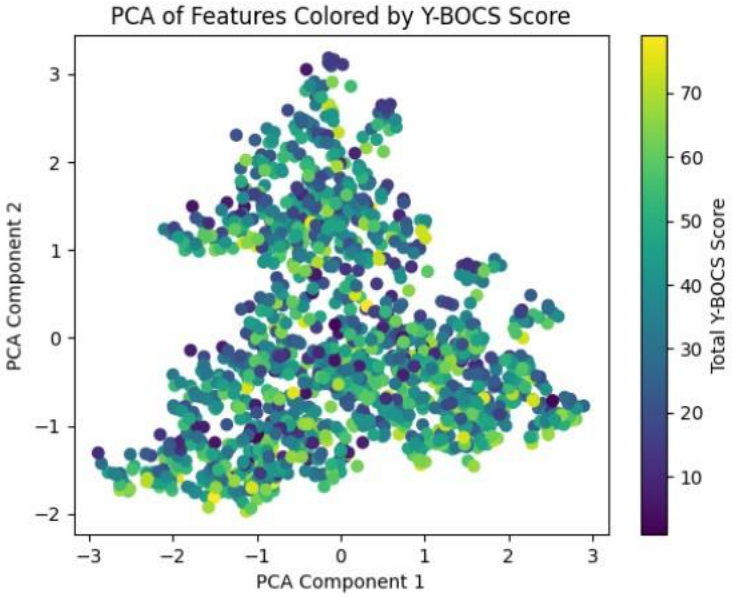


Correlation heatmap reveals key relationships among the numerical features in the dataset. The Total Y-BOCS Score shows a strong positive correlation with both Y-BOCS Obsessions (0.72) and Compulsions (0.72) , indicating that higher scores in these subscales are closely associated with greater overall symptom severity. In contrast, age displays negligible correlations with other variables (values near zero), suggesting it has little linear relationship with symptom duration or severity. The duration of symptoms also exhibits weak correlations , with the strongest being a slight negative association with compulsions (-0.04). These findings emphasize that obsession and compulsion scores are strongly interrelated with the Y-BOCS assessment, while demographic factors like age and symptom duration appear less influential.

Interaction of categorical variable with numerical variable- PCA.

We encode categorical variables like gender,ethnicity,obsession type,compulsion using one-hot encoding . These encoded variables were combined with numerical features (Age,Duration of

Symptoms(months), Total Y-BOCS Score) to form the final feature matrix.All features were standardized using StandardScaler to ensure uniform scaling which is essential for PCA. PCA was applied to reduce the feature space to two princiap components (n_components = 2). These components capture the maximum variance in the data while simplifying the visualization.



The Principal component analysis scatter plot illustrates the distribution of samples based on dimensionality reduction , with points coloured by Total Y-BOCS Score. The visualization reveals distinct clustering patterns, where samples with higher Y-BOCS scores (indicating greater symptom severity) tend to group in specific regions, while lower-scoring cases are more dispersed. This suggests that the principal components capture meaningful variance related to symptom severity. A few outliers , particularly near extreme values of PC1 and PC2 , may represent atypical cases requiring further investigation. While the plot highlights potential relationships between features and symptom severity , interpreting the exact influence of each variable would require examining component loadings. This analysis supports the utility of PCA for identifying underlying patterns in data , which could inform subsequent modelling or subgroup analyses.

Machine Learning Pipeline for Y-BOCS Score Prediction

We implement an end-to-end machine learning pipeline to predict the Total Y-BOCS Score (a measure of OCD symptom severity). The details of the pipeline are as follows.

1. Data Pre-processing & Feature Engineering

- **One-Hot Encoding**:
  Convert categorical variables (eg. Gender, ethnicity) into numerical format using OneHotEncoder, dropping the first category to avoid multicollinearity.
- **Feature Engineering**:
  Creat interaction terms (eg. Age*duration of symptoms) and binned continuous variables (eg. Duration,binned) to capture non-linear relationships.
- **Standardization**:
  Scale numerical features (eg. Age, Y-BOCS sub score) using Standardscaler to ensure equal contribution to the model.

2. Feature Selection

- **Recursive Feature Elimination (RFE)**:
  Use Random forest regressor to select the top 10 most predictive features, reducing dimensionality and improving model interpretability.

3. Model Training & Evaluation

- **Pipeline Construction**:
  Combine pre-processing and modelling (XGBRegressor) into a single pipeline for reproducibility.
- **Performance Metrics**:
  Evaluate the model using R^2 score and Mean Absolute Error on the test set, indicating how well the model explains variance in Y-BOCS scores.
- **Actual vs. Predicted Plot**:
  Visualize model predictions against true values to check for systematic biases.

4. Model Interpretability

- **SHAP Analysis**:
  SHAP values explain feature importance, identifying key drivers (eg. Y_BOCS Obsessions, Duration of Symptoms) of predicted scores.
- **Feature Importance Plot**:
  Rank top 10 influential features using RandomForestRegeressor.

5. Patient Segmentation

- **K-Means Clustering**:

  Grouped patients into 3 clusters based on clinical features to uncover subtypes (eg. "High Severity/Long Duration").

- **Cluster Visualization**:

  Plotted clusters by age and duration of symptoms to reveal distinct patient profiles.

6. Deployment Readiness

Save Models:

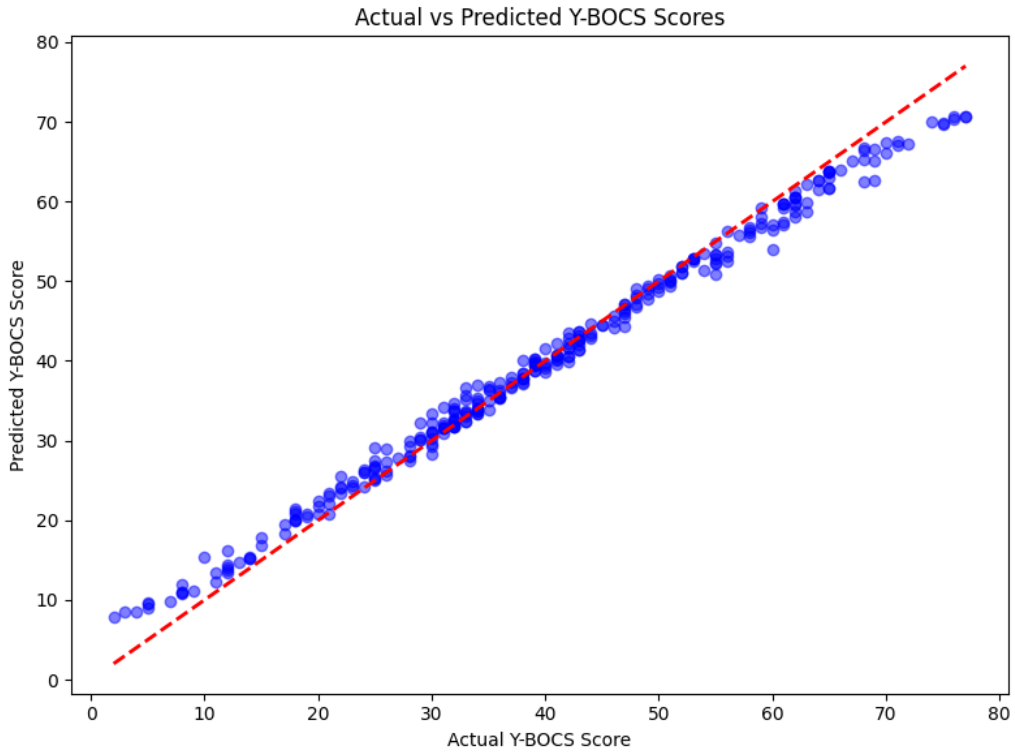
Export the trained pipeline(ocd_model.pkl) and clustering model(kmeans_model.pkl)for future use.

- **Saved Visualizations**:

  Stored plots(shap_plot.png, cluster_plot.png)) for reporting and documentation.

Plot explanations

1. **Actual vs. Predicted Y-BOCS Scores Plot**



This graph compares the models predicted Y-BOCS scores (vertical axis) to the actual scores reported

by patients (horizontal axis). The closer the points are to the red diagonal line, the more accurate the predictions. This shows how well the model estimates symptom severity, helping clinicians identify cases where predictions might need adjustment.

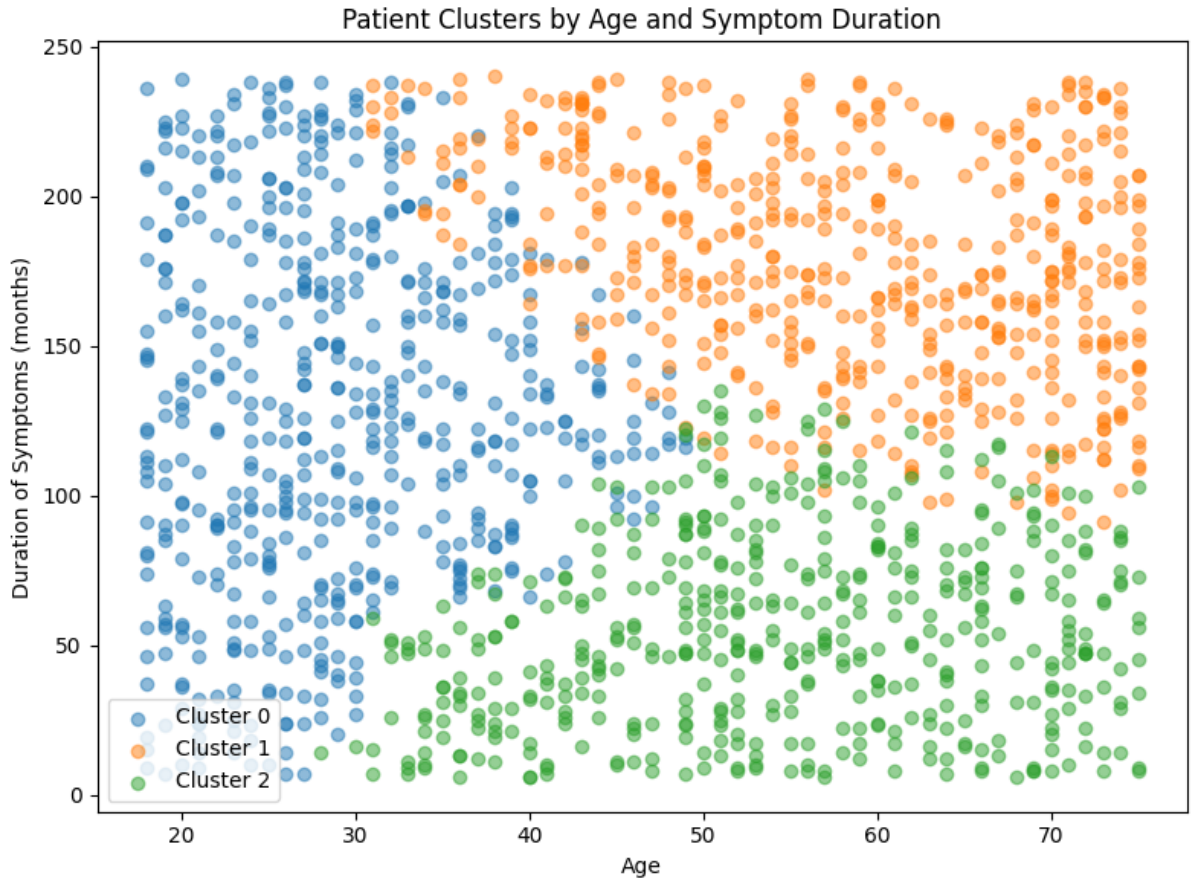
2. **SHAP Feature Importance Plot**

This chart highlights which factors most influence the models predictions. Features like Y-BOCS Obsessions/Compulsions (top of the list) have the biggest impact.Higher values increase predicted

severity. Demographic factors (e.g ethnicity) and treatment history (e.g medications) also play smaller but notable roles. The longer the bar, the stronger the effect.

3. **Patient Clusters by Age and Symptom Duration**

```
Cluster Distribution:
Cluster
0    531
1    502
2    467
Name: count, dtype: int64
Cluster Statistics:
         Total Y-BOCS Score        Age  Duration of Symptoms (months)
\
Cluster
0                 38.962335  29.020716                     127.935970
1                 40.511952  58.424303                     178.260956
2                 39.582441  54.460385                      53.955032

         Y-BOCS Score (Obsessions)  Y-BOCS Score (Compulsions)
Cluster
0                        19.807910                   19.154426
1                        20.314741                   20.197211

2                        20.034261                   19.548180
```

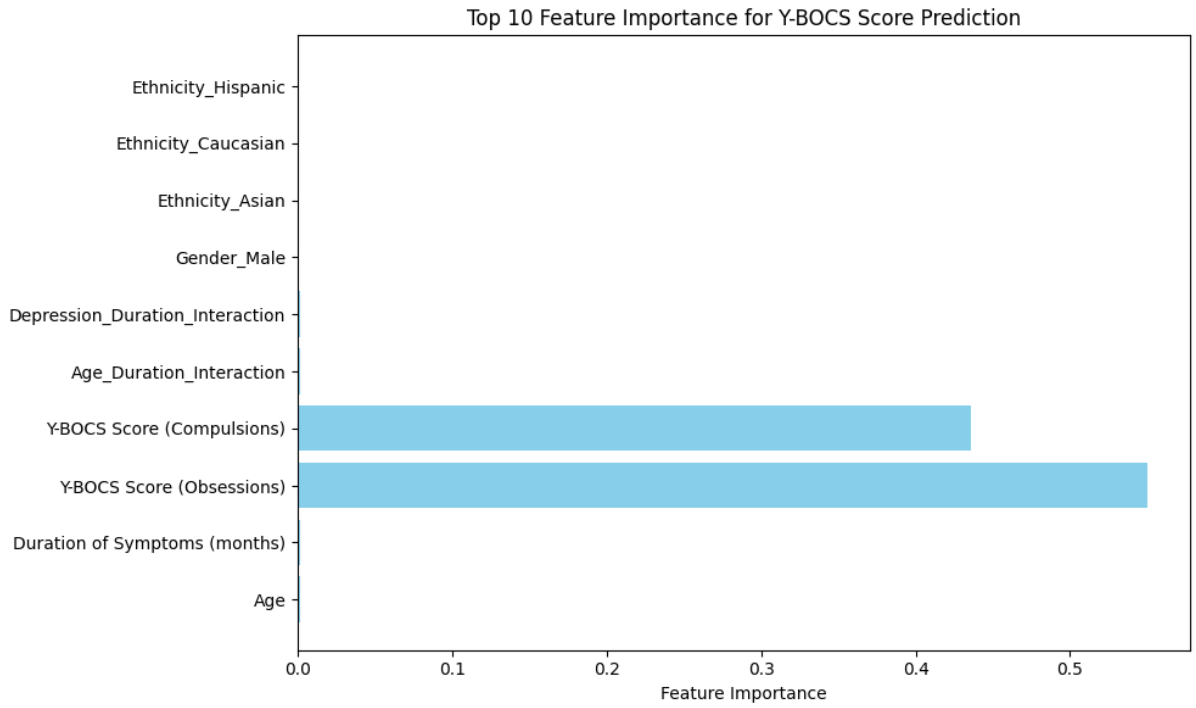


Here, patients are grouped into three clusters based on age and how long they've had symptoms.

- **Cluster 1 (Blue):** Younger patients with shorter symptom duration.
- **Cluster 2 (Orange):** Middle-aged patients with moderate duration.
- **Cluster 3 (Green):** Older patients with long-standing symptoms.

  This helps tailor treatments to different patient profiles.

4. **Top 10 Feature Importance Plot**



This bar chart ranks the most important predictors of Y-BOCS scores. Obsessions/Compulsions scores and symptom duration are the strongest drivers, while demographics (eg. ethnicity) and interaction terms (eg. age × duration) contribute less but still matter. The taller the bar, the more the feature affects predictions.

Summary

These results show that OCD symptom severity (Y-BOCS scores) can be predicted using patient data, with obsessions/compulsions being the strongest predictors. The models accuracy is

`R² Score (Test): 0.9847`

`MAE (Test): 1.6054.`

SHAP and feature importance plots reveal why it makes certain predictions. Clustering further helps categorize patients into subgroups for personalized care. Together, these tools provide actionable insights for clinicians and researchers.

Conclusion

The data science project on OCD patient dataset provides a comprehensive analysis of demographic and clinical factors influencing OCD symptom severity, as measured by the Y-BOCS scores. Through exploratory data analysis (EDA), we uncovered key insights , such as the balanced gender distribution , diverse age groups , and the strong correlation between obsession and compulsion sub-scores. Feature engineering , including the creation of a Total Y-BOCS Score and age binning , enhanced the dataset for further modelling.

The machine learning pipeline successfully predicted Y-BOCS scores with high accuracy ($R^2 = 0.9847$, MAE = 1.6054) , demonstrating the effectiveness of XGBoost and feature selection techniques like RFE. SHAP analysis highlighted that Y-BOCS sub-scores and symptom duration were the most influential predictors, while demographic factors played a smaller but notable role. Additionally, K-means clustering identified distinct patient subgroups, enabling personalized treatment strategies based on age and symptom duration.

This project not only advanced our understanding of OCD severity drivers but also provided actionable tools for clinicians, such as predictive models and patient segmentation. Future work could explore longitudinal data to track symptom progression and incorporate additional clinical variables for even more refined predictions. Overall, the findings underscore the potential of data science in improving mental health diagnostics and treatment planning.