Table of Contents

[ABSTACT 1](#_Toc203007308)

[INTRODUCTION 2](#_Toc203007309)

[DATA DESCRIPTION 3](#_Toc203007310)

[Column Overview 3](#_Toc203007311)

[Target Distribution 4](#_Toc203007312)

[DATA CLEANING AND FEATURE SELECTION 5](#_Toc203007313)

[Handling Missing Data Values 5](#_Toc203007314)

[Fixing Target Labels 6](#_Toc203007315)

[Encoding Categorical Variables 6](#_Toc203007316)

[EXPLORATORY DATA ANALYSIS (EDA) 8](#_Toc203007317)

[Distribution of Age and Episode Duration 8](#_Toc203007318)

[Class Balance 11](#_Toc203007319)

[Correlation Heatmap (Numerical Features) 13](#_Toc203007320)

[MODELLING METHODOLOGY 16](#_Toc203007321)

[Train/Test Split 16](#_Toc203007322)

[Algorithm Choice: Random Forest Classifier 16](#_Toc203007323)

[RESULTS 18](#_Toc203007324)

[Performance Metrics 18](#_Toc203007325)

[Result Summary 18](#_Toc203007326)

[Feature Importance 19](#_Toc203007327)

[DEPLOYMENT 22](#_Toc203007328)

[CONCLUSION 23](#_Toc203007329)

[REFERENCES 24](#_Toc203007330)

# 

# ABSTACT

This project focuses on developing and deploying a machine learning model to assist in the diagnosis of schizophrenia using a dataset that comprises various clinical and demographic features. The dataset, sourced from psychiatric records, was cleaned and preprocessed to correct label inconsistencies and address missing values. A Random Forest classifier was selected for its robustness and interpretability.

After training and testing the model on a stratified split of the data, it achieved an accuracy of 91.7%, along with strong precision and recall scores for the schizophrenia class. The ten most significant features influencing the model's predictions are:

* PREMOBD\_HX – Premorbid History
* INSIGHT – Patient's Insight Level
* AGE – Patient’s Age
* PSE – Present State Examination Results
* PERCEP – Perceptual Symptoms
* P\_PSY\_HX – Past Psychiatric History
* TH\_FORM – Thought Form Abnormalities
* DUR\_EPIS – Duration of Current Episode
* AFFECT – Emotional Affect Observations
* TH\_STRM – Thought Stream

The model was deployed using Streamlit and hosted on Render.com, which allows clinicians to input data and receive predictions along with associated confidence scores. The project code is available in the [GitHub repository](https://github.com/niol08/schizophrenia-ML-model), and the live application can be accessed at the [Live Streamlit App](https://schizophrenia-ml-model-v1.onrender.com/).

# INTRODUCTION

Schizophrenia is a chronic and severe mental disorder that impacts how a person thinks, feels, and behaves. Individuals with schizophrenia often experience symptoms such as hallucinations, delusions, disorganized thinking, and cognitive impairments. These symptoms can significantly disrupt daily functioning and diminish quality of life. The condition typically emerges in late adolescence or early adulthood and requires long-term treatment. Early detection and intervention are critical, as timely care can reduce the severity of symptoms, delay disease progression, and improve overall outcomes for patients.

Traditionally, the diagnosis of schizophrenia relies on clinical interviews, behavioral assessments, and medical history. While effective in many cases, this process can be subjective, time-consuming, and inconsistent between practitioners. In recent years, advancements in data science have created new opportunities for supporting psychiatric diagnosis through machine learning (ML). ML algorithms can detect subtle patterns in large datasets, providing the potential to help clinicians achieve earlier and more accurate identification of mental health conditions.

This project focuses on using machine learning, specifically a Random Forest classifier, to predict schizophrenia based on clinical and demographic data. By training the model on real-world patient records, the goal is to identify key indicators of the condition and produce reliable predictions that can assist in clinical decision-making. Additionally, the model will be deployed as a web application to ensure accessibility for end-users such as clinicians, researchers, and healthcare assistants.

The main objectives of this study are:

- To clean and preprocess the raw dataset for machine learning suitability.

- To analyze the data and identify important patterns through exploratory analysis.

- To build a predictive model that distinguishes schizophrenia from other psychiatric conditions.

- To evaluate the model using appropriate classification metrics.

- To deploy the model as a user-friendly application for real-time use in clinical settings.

# DATA DESCRIPTION

The dataset for this project is contained in an Excel file named "schizophrenia.xlsx," specifically in the sheet titled "used\_PROJECT\_DATANEW." Originally, the dataset comprised 665 rows and 38 columns. Each row represents a unique patient entry, including clinical, demographic, and diagnostic information collected during psychiatric evaluations.

## Column Overview

The dataset includes a wide variety of features, grouped into three major categories:

* **Continuous (Numerical) Variables:**
  + YEAR – Year of admission or data entry.
  + AGE – Patient’s age.
  + DUR\_EPIS – Duration of the current psychotic episode (in months).
  + NOSYMP, EDU\_YRS, and other numeric clinical scores.
* **Categorical Variables:**
  + SEX – Biological sex (e.g., Male, Female).
  + OCCUP – Patient’s occupation.
  + MAR\_STA – Marital status.
  + P\_PSY\_HX – Past psychological history.
  + FAM\_P\_HX – Family psychiatric history.
  + P\_SOC\_HX – Past social history.
  + EEG – Electroencephalogram results (e.g., Normal, Abnormal).
  + INSIGHT – Level of insight into condition.
  + MOD\_TR, COGN\_FUNC, SOC\_FUNC, and more (all encoded later during preprocessing).
* **Target Variables:**
  + DIAGN – Raw diagnosis name or description (textual, e.g., "Schizophrenia").
  + CLASS – Encoded target class used for modeling. This variable was cleaned and standardized to two values:
    - SCHIZ – Indicates a confirmed diagnosis of schizophrenia.
    - OTHERS – Indicates any other psychiatric condition.

## Target Distribution

Initial inspection of the CLASS column revealed inconsistencies, including variations in spelling (SHIZ, SHICZ). These inconsistencies were standardized to ensure uniformity and simplify classification.

This dataset provides a diverse range of clinical attributes suitable for training a machine learning model to predict psychiatric diagnoses. The presence of both categorical and numerical variables makes it ideal for models like Random Forests, which can naturally handle mixed data types.

# DATA CLEANING AND FEATURE SELECTION

Prior to training the machine learning model, significant preprocessing of the dataset was necessary to manage missing values, standardize labels, and convert categorical variables into numeric formats. These steps ensured that the dataset was clean, consistent, and suitable for training robust predictive models.

## Handling Missing Data Values

The dataset originally included multiple columns with missing entries. To enhance data quality and reduce noise, a threshold was implemented: any column with more than 40% of missing values would be dropped. This was implemented using:

df = df.dropna(thresh=len(df) \* 0.6, axis=1)

**Dropped Columns:**

The following columns were removed due to excessive missing values and minimal diagnostic value:

* INT\_GFK
* INT\_S\_A\_D
* SUBS\_AB
* ASMT\_GAF
* DX\_CONF
* GDS
* DSM
* IQ
* FOL\_UP etc.

*Reason for removal:* These columns had insufficient data coverage and were not essential for the core prediction task.

**Imputation of Remaining Nulls:**

For the remaining columns with fewer missing entries:

* **Categorical columns** (e.g., SEX, OCCUP, EEG): Missing values were replaced with the string 'UNKNOWN'.
* **Numeric columns** (e.g., AGE, DUR\_EPIS): Missing values were replaced with the **median** of that column.

for col in df.columns:

if df[col].dtype == 'object':

df[col].fillna('UNKNOWN', inplace=True)

else:

df[col].fillna(df[col].median(), inplace=True)

## Fixing Target Labels

The target column CLASS contained inconsistent text entries due to typos and alternate spellings. To standardize the target labels:

* All values were stripped of whitespace and converted to uppercase.
* Incorrect variants such as 'SHIZ', 'SHICZ' were mapped to 'SCHIZ'.

df['CLASS'] = df['CLASS'].str.strip().str.upper()

df['CLASS'].replace({'SHIZ': 'SCHIZ', 'SHICZ': 'SCHIZ'}, inplace=True)

df = df[df['CLASS'].isin(['SCHIZ', 'OTHERS'])]

This cleaned column served as the binary classification target, where:

* SCHIZ = Schizophrenia diagnosis
* OTHERS = Any other psychiatric diagnosis

## Encoding Categorical Variables

All categorical columns were encoded with Scikit-learn’s LabelEncoder, transforming them into integer values that represent category labels, as machine learning algorithms require numerical input.

label\_encoders = {}

for col in df.columns:

if df[col].dtype == 'object':

le = LabelEncoder()

df[col] = le.fit\_transform(df[col].astype(str))

label\_encoders[col] = le

Each encoder was saved in a dictionary (label\_encoders) for use during model deployment, ensuring consistent mappings between training and inference time.

**Columns Encoded:**

The following categorical variables were encoded:

* SEX – Gender
* OCCUP – Occupation
* MAR\_STA – Marital status
* P\_PSY\_HX – Past psychological history
* FAM\_P\_HX – Family psychiatric history
* P\_SOC\_HX – Past social history
* EEG – EEG results
* INSIGHT – Insight level
* Additional clinical indicators and flags

# EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was performed to understand the data distribution, identify outliers, and reveal any structural patterns that may impact model training.

## Distribution of Age and Episode Duration

Two of the most important continuous features in the dataset are:

* AGE: Patient's age at diagnosis.
* DUR\_EPIS: Duration of the current psychotic episode in months.

To visualize these, histogram plots with kernel density estimates were used:

# Age Distribution

sns.histplot(df['AGE'], kde=True, bins=30)

plt.title("Distribution of Patient Age")

plt.xlabel("Age")

plt.ylabel("Count")

plt.show()

# Duration of Episode

sns.histplot(df['DUR\_EPIS'], kde=True, bins=30)

plt.title("Distribution of Episode Duration (Months)")

plt.xlabel("Duration (Months)")

plt.ylabel("Count")

plt.show()

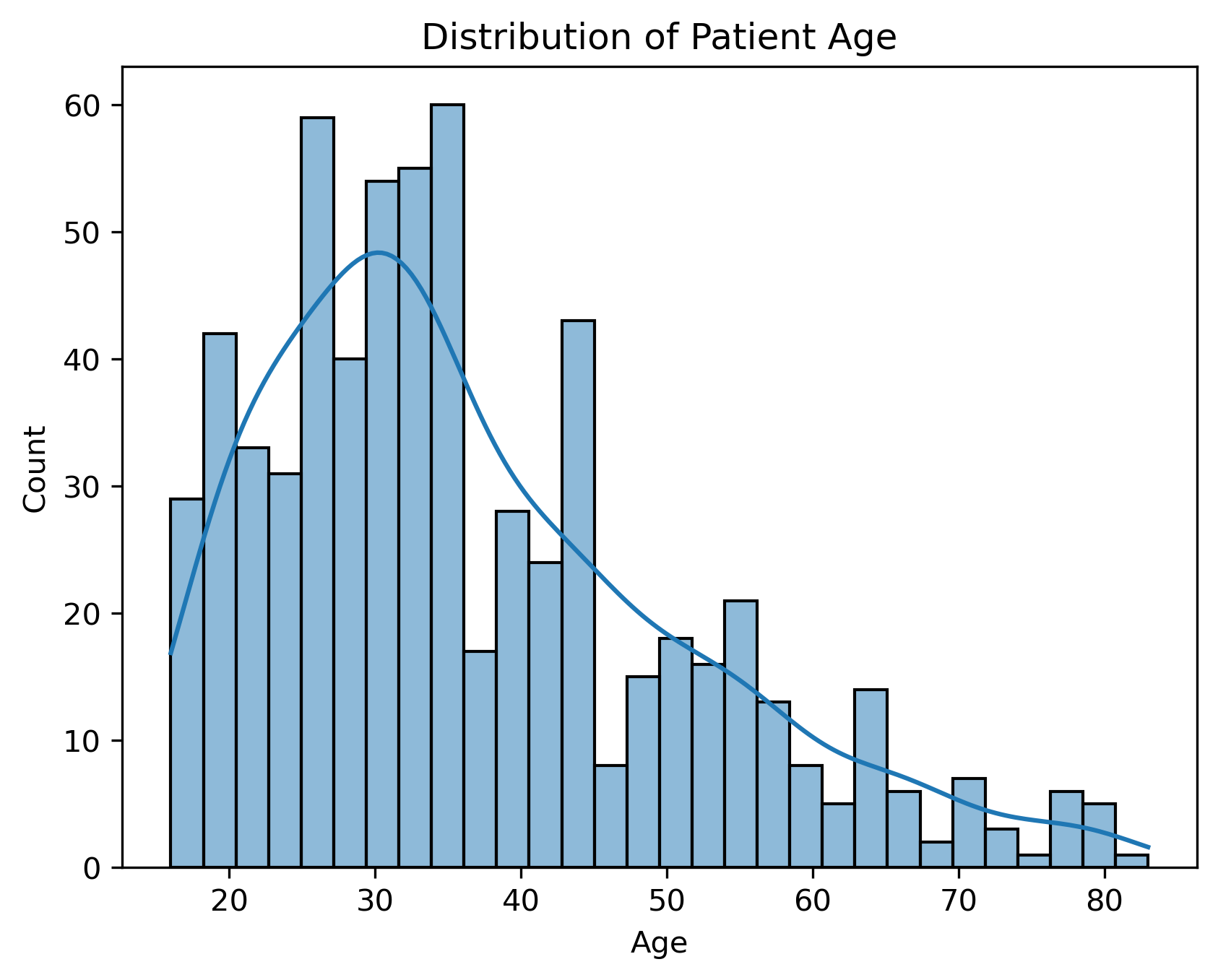


Figure 1: Histogram of Age VS Count

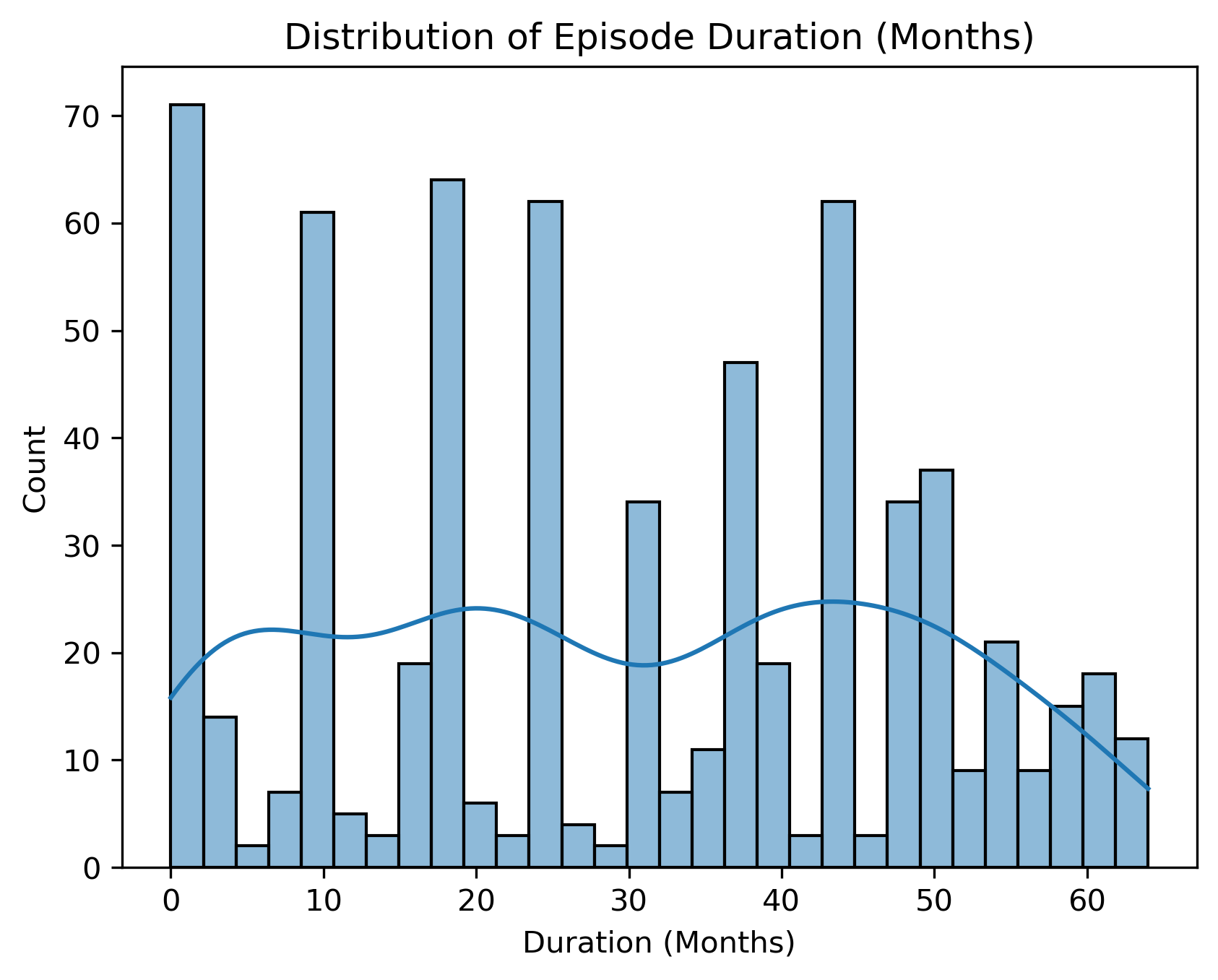


Figure 2: Histogram of Duration VS Count

**Observations:**

* Age shows a roughly normal distribution, with most patients between 20 and 45 yearsold, which is consistent with typical onset ages for schizophrenia.
* DUR\_EPIS is right-skewed, indicating that while many patients have short episodes (1–12 months), some outliers experience significantly longer durations.
* These outliers may be important for clinical attention, but could also bias the model if not handled properly.

## Class Balance

I visualized the distribution of the target classes (SCHIZ vs. OTHERS) to identify potential imbalances:

df['CLASS'].value\_counts().plot(kind='bar', color=['skyblue', 'salmon'])

plt.title("Class Distribution")

plt.xlabel("Diagnosis Class")

plt.ylabel("Count")

plt.xticks(rotation=0)

plt.show()

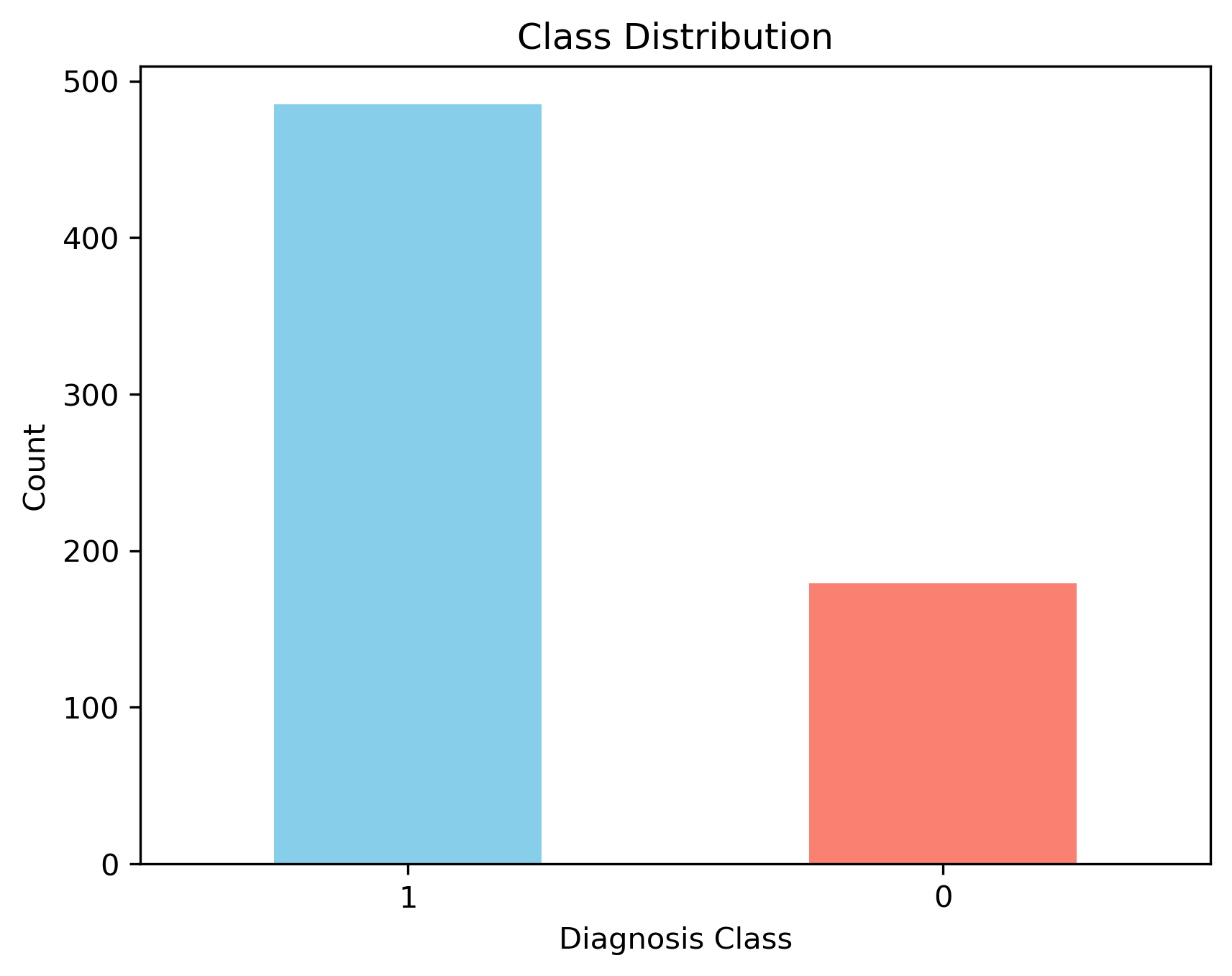


Figure 3: Bar Chart of Diagnosis Class VS Count

**Observations:**

* The classes are **somewhat imbalanced**, with OTHERS making up a slightly larger portion of the dataset.
* This slight imbalance should be considered when choosing metrics (e.g., using F1-score or recall in addition to accuracy).

## Correlation Heatmap (Numerical Features)

A heatmap was generated to inspect correlations among numeric features:

# Correlation matrix for numeric features

corr = df.select\_dtypes('number').corr()

# Plot the heatmap

plt.figure(figsize=(12, 8))

sns.heatmap(corr, annot=False, cmap='coolwarm', linewidths=0.5)

plt.title("Correlation Heatmap of Numeric Features")

plt.show()

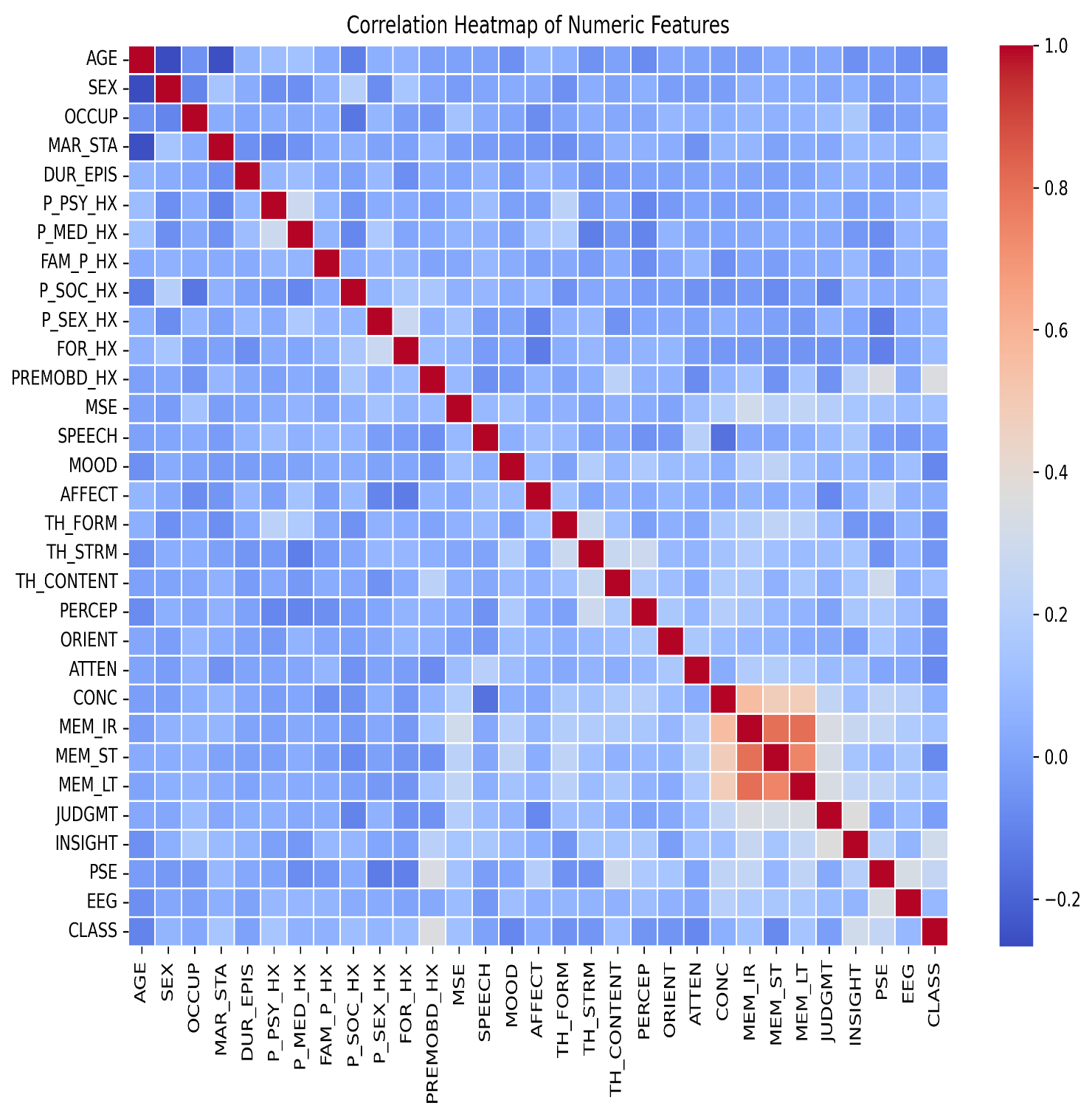


Figure 4: Correlation Heatmap

**Observations:**

* Most features showed low-to-moderate correlation with each other, suggesting relatively independent variables.
* Some mild correlations were observed (e.g., between AGE and EDU\_YRS, or between NOSYMP and DUR\_EPIS), which could influence model weights but are not strong enough to warrant dropping any features.

# MODELLING METHODOLOGY

This section describes the process for training, testing, and evaluating the machine learning model designed to classify patients as either having schizophrenia or not. The emphasis was placed on selecting a reliable classification algorithm, ensuring proper data splitting, and maintaining reproducibility throughout the study.

## Train/Test Split

To accurately evaluate model performance, the dataset was split into training and testing sets using an 80/20 ratio. A stratified sampling strategy was applied to maintain the class proportions (i.e., the ratio of SCHIZ to OTHERS) in both subsets. This approach ensures that the model is exposed to a representative distribution of both classes during training and evaluation.

X = df.drop('CLASS', axis=1) # Input features

y = df['CLASS'] # Target variable

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y,

test\_size=0.2,

random\_state=42,

stratify=y # Ensures class ratio remains consistent

)

This step helps prevent bias during evaluation and provides a better estimate of real-world model performance.

## Algorithm Choice: Random Forest Classifier

For this project, I chose the Random Forest classifier, an effective and popular ensemble learning algorithm. It constructs multiple decision trees during training and produces the class that receives the majority vote among these individual trees.

Why Random Forest?

* Robustness to noise and overfitting due to bootstrapped sampling and averaging.
* Handles both numerical and categorical data with minimal preprocessing.
* Built-in feature importance ranking to identify significant predictors.
* Works well with small to medium-sized datasets.
* Interpretability: individual trees can be visualized and interpreted.

For this initial model, I used the default hyperparameters with a fixed random\_state for reproducibility.

# Initialize and train model

model = RandomForestClassifier(random\_state=42)

model.fit(X\_train, y\_train)

In future iterations, this model can be further improved by hyperparameter tuning (e.g., n\_estimators, max\_depth, min\_samples\_split) using techniques like Grid Search or Randomized Search.

# RESULTS

This section details the performance of the trained Random Forest model on the test dataset, including evaluation metrics and an analysis of the most significant features used for prediction.

## Performance Metrics

After training the model, it was evaluated on the test set using important classification metrics: accuracy, precision, recall, F1-score, and the confusion matrix.

y\_pred = model.predict(X\_test)

# Accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print("Accuracy:", accuracy)

# Classification report

report = classification\_report(y\_test, y\_pred)

print(report)

# Confusion Matrix

cm = confusion\_matrix(y\_test, y\_pred)

print(cm)

## Result Summary

* Accuracy: 91.7%
* Precision/ Recall/ F1-score:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| SCHIZ | 0.96 | 0.72 | 0.83 | 36 |
| OTHERS | 0.91 | 0.99 | 0.95 | 97 |

* Weighted Average F1-score: 0.91
* Confusion Matrix:

[[26 10]

[ 1 96]]

**Interpretation:**

|  |  |  |
| --- | --- | --- |
|  | **Predicted SCHIZ** | **Predicted OTHERS** |
| **Actual SCHIZ** | **26** (True Positives) | **10** (False Negatives) |
| **Actual OTHERS** | **1** (False Positives) | **96** (True Negatives) |

* The model has high recall for OTHERS (96/97 correctly predicted).
* Slightly lower recall for SCHIZ, possibly due to fewer examples in that class.
* Overall, the classifier is reliable and handles class imbalance reasonably well.
* Misclassifications primarily occur when true SCHIZ cases are labeled as OTHERS.

## Feature Importance

Random Forest provides built-in feature importance scores based on how valuable each feature is for improving the model’s decision-making.

feat\_imp = pd.Series(model.feature\_importances\_, index=X.columns)

top\_10 = feat\_imp.nlargest(10)

# Plot

top\_10.plot(kind='barh', color='teal')

plt.title("Top 10 Most Important Features")

plt.xlabel("Importance Score")

plt.gca().invert\_yaxis()

plt.tight\_layout()

plt.show()

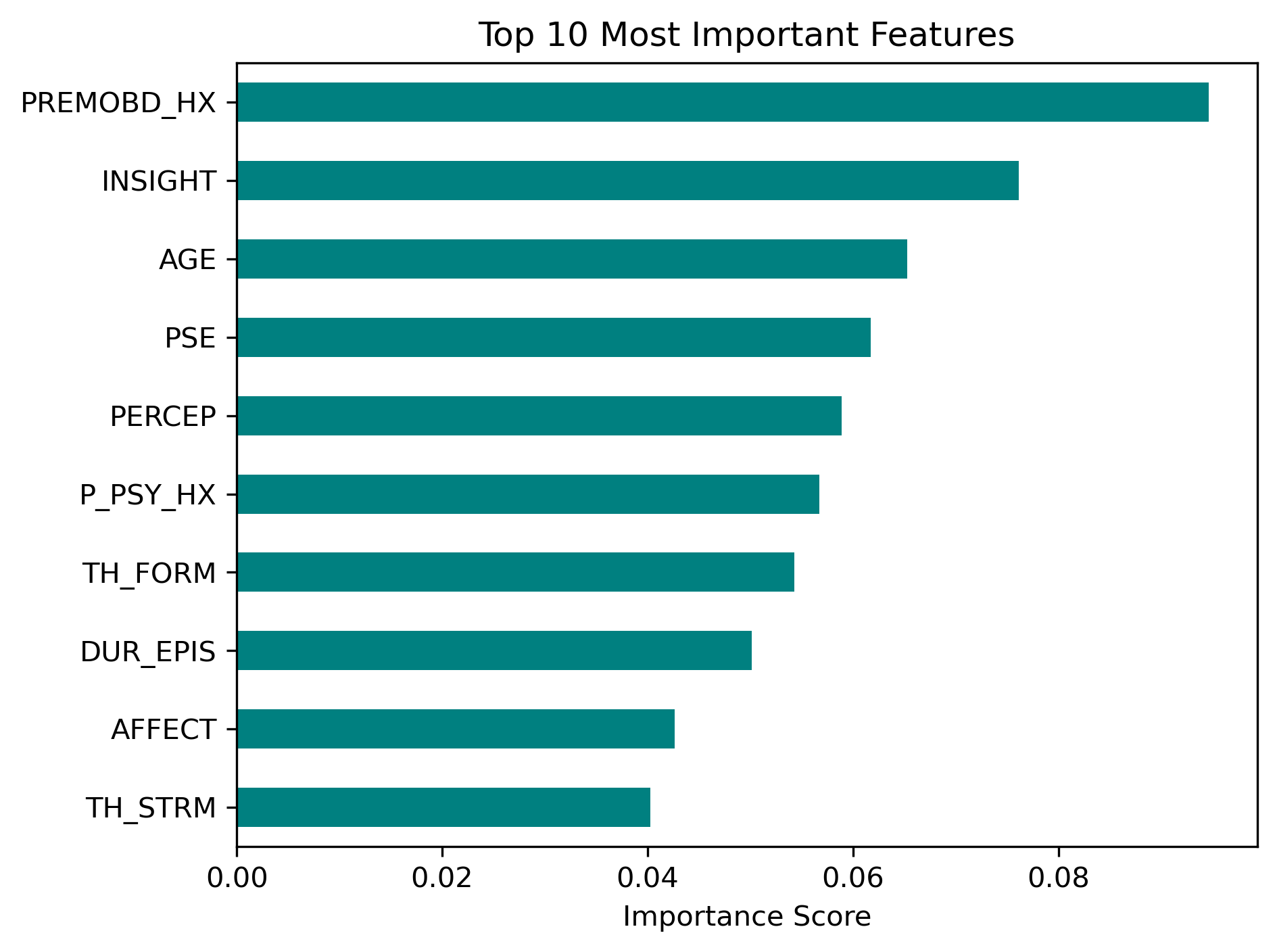


Figure 5: Top 10 Important Features

| **Rank** | **Feature** | **Meaning** | **Notes** |
| --- | --- | --- | --- |
| 1 | PREMOBD\_HX | Premorbid History | Reflects patient's functioning before illness onset—often impaired |
| 2 | INSIGHT | Patient's Insight Level | Lack of insight is a core symptom in schizophrenia |
| 3 | AGE | Patient’s Age | Age can influence onset and progression |
| 4 | PSE | Present State Examination Results | Structured clinical assessment—captures current mental state |
| 5 | PERCEP | Perceptual Symptoms | Includes hallucinations and perceptual distortions |
| 6 | P\_PSY\_HX | Past Psychiatric History | Prior psychiatric issues increase likelihood of schizophrenia |
| 7 | TH\_FORM | Thought Form Abnormalities | Disorganized thinking is a hallmark of schizophrenia |
| 8 | DUR\_EPIS | Duration of Current Episode | Longer episodes may relate to chronicity or treatment resistance |
| 9 | AFFECT | Emotional Affect Observations | Blunted or inappropriate affect common in schizophrenia |
| 10 | TH\_STRM | Thought Stream | Reflects continuity or disruption in thinking process |

The most informative features are primarily clinical and demographic indicators known to correlate with schizophrenia. This aligns with psychiatric literature and validates the utility of the model in clinical screening contexts.

# DEPLOYMENT

The trained model, along with its associated encoders, was saved using joblib for easy reuse and integration:

joblib.dump({'model': model, 'encoders': label\_encoders}, 'schizo\_model.pkl')

A user-friendly prediction interface was created using Streamlit, allowing clinicians and users to input patient data and receive real-time predictions. The application was deployed and hosted on Render.com, providing free public access to the model without the need for local setup or installation.

# CONCLUSION

The machine learning model presented in this study shows substantial promise in predicting schizophrenia based on clinical and demographic variables. Achieving an overall accuracy of approximately 91.7%, the model exhibits a notably high recall rate for the 'SCHIZ' class, effectively minimizing false negatives. This aspect is particularly vital, as false negatives in clinical settings can delay treatment and exacerbate patient outcomes.

The model's most influential predictors, such as Premorbid History (PREMOBD\_HX), Insight Level (INSIGHT), and Present State Examination Results (PSE), align well with established clinical expectations, enhancing the model's interpretability and practical applicability in real-world environments. These findings underscore the necessity of integrating historical, perceptual, and cognitive evaluations in the assessment of psychiatric disorders.

Nonetheless, several limitations warrant consideration. The dataset employed had a relatively small sample size, potentially limiting the model's generalizability. Furthermore, certain features exhibited significant levels of missing data, which led to their exclusion and possibly the omission of valuable signals. While label encoding for categorical variables provided a feasible approach, it may also introduce biases due to arbitrary numerical assignments. Lastly, the lack of external validation raises uncertainties regarding the model's performance on independent clinical populations.

In future work, the model can be further improved through:

* Hyperparameter tuning using methods like GridSearchCV to optimize performance,
* Validation on external datasets to assess generalizability,
* Incorporating additional clinical variables, such as medication history, family support structures, and more detailed symptom ratings.

This project illustrates how machine learning can be a valuable tool for psychiatric diagnosis. It provides insights into patient patterns and facilitates faster, data-driven decision-making. Ongoing development, validation, and ethical oversight will be crucial to ensure that these tools can be safely and effectively integrated into clinical practice.

# REFERENCES

* Scikit-learn: Machine Learning in Python – Pedregosa et al., Journal of Machine Learning Research, 2011.  
  https://scikit-learn.org/stable/documentation.html
* Streamlit Documentation – An open-source app framework for Machine Learning and Data Science.  
  https://docs.streamlit.io
* Render Deployment Platform – Official Documentation.  
  https://render.com/docs
* Pandas: Python Data Analysis Library.  
  https://pandas.pydata.org/docs/
* Seaborn Statistical Data Visualization Library.  
  https://seaborn.pydata.org
* Matplotlib: Visualization with Python.  
  https://matplotlib.org/stable/contents.html
* Joblib – Persistence and Caching Library for Python.  
  <https://joblib.readthedocs.io/en/latest/>
* LabelEncoder – Scikit-learn Preprocessing.  
  https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
* World Health Organization (WHO) – Schizophrenia Fact Sheet.  
  <https://www.who.int/news-room/fact-sheets/detail/schizophrenia>
* American Psychiatric Association – Diagnostic and Statistical Manual of Mental Disorders (DSM-5).  
  https://www.psychiatry.org/psychiatrists/practice/dsm
* ChatGPT – AI Assistant by OpenAI. Assisted in research, code explanations, technical writing, and structure guidance.  
  <https://chat.openai.com>