International Journal of Allied Practice, Research and Review

# Analysis and prediction on Mental Health Disorder

**P.S Thakur,**
**Research Scholar, Department of Computer Science,**
**OPJS University, Churu, Rajasthan, India**

**Abstract: - Analysis and prediction of mental health disorders is an important area of research as mental health issues affect a large number of people worldwide. Machine learning algorithms can be used to identify patterns in large datasets of mental health records and predict the likelihood of developing a mental health disorder.**

**In this study, we aim to develop a machine learning model to predict mental health disorders based on various demographic and clinical variables. We collected a dataset of mental health records from multiple sources and preprocessed the data to remove missing values and outliers. We then trained and evaluated multiple machines learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks, to identify the best performing algorithm for the prediction of mental health disorders.**

**Our results indicate that the random forest algorithm outperformed the other algorithms, achieving an accuracy of 85%. The most important predictors in the model were age, gender, employment status, and previous history of mental health disorders.**

**This study has implications for the development of predictive models for mental health disorders that can aid in early identification and treatment of these disorders. Future studies could focus on incorporating additional variables, such as genetic factors and social determinants of health, to improve the accuracy of these models.**

**Keywords: - Disorders, NDA, Machine learning algorithms, Deep learning algorithm.**

## I. Introduction

With the advent of computers and information technology in the health services, amount of data has increased rapidly for various kinds of illnesses. These data sets can be obtained from health centers, where data of patients are maintained in customized softwares. Such data sets are also available on the online platforms for researchers. In the modern world Mental Illnesses are common type of sufferings that are on the rise these days. Persons belonging to any age group can suffer from mental Illness and Disorders. Numbers of Medicos qualified to treat such disorders are far lesser as compared to the number of patients. Diagnosis involves the discussion with the patient. A questionnaire about the symptoms, medical history, genetic background etc. is discussed with the patient. Psychological tests are performed on the patients to advice a treatment. Data so gathered from the patients is either in certain form or uncertain form. Certain data is easy to operate upon but on uncertain data Doctors have to apply probability factors to treat the patients. Now a days, a branch of artificial intelligence deals with intelligent systems.

These systems have customized a machine learning algorithm that deals with logical deductions. To operate with any ML algorithm, the first step is data analysis. Data analysis involves data cleaning, data transforming, data modeling and finally developing conclusion in order to find the results. Exploratory data analysis (EDA) is a process to analyze the data sets in order to represent their main characteristics using statistical graphics and data visualization techniques. It helps to formulate hypothesis that could leads to form more data experimentation. So, initial data analysis technique is converted to form EDA. Many popular languages are available to perform data analysis, Python is one of them. Python open-source programming language is used for data mining and machine learning. It has many libraries to represent data in the form of charts, tables, graphs and plots. The libraries like Numpy and Pandas makes it easy for a researcher to operate on mathematical and statistical data sets.The libraries we are going to use in this paper are------.

The paper is divided into four sections. In the Section-1, a brief introduction about the research to be implemented on the psychological characteristics of a patient is given. The paper works on data sets available in six mental health illnesses predominant among adults. These illnesses are Schizophrenia, bipolar disorder, eating disorder, anxiety disorder, alcohol disorder and drug used disorder. the author performs exploratory data analysis on the open source data sets taken from Kaggle containing global trends in mental health disorder. In which age depression is more common? Are men or women more likely to have depression?

## II. Literature Review

Ceyhun Ozgur et.al, 2015 has given a description about the voluminous data and focused on how any statistical software be choose to solve a big data problem. Gyeongcheol Cho et.al, 2018 described about various machine learning algorithms used by the researchers to solve the mental illness. Dr. Ayesha Bhanu et.al explained the exploratory data analysis and its various techniques in elaborative form with a implementation on employee attrition system.

In recent years, machine learning algorithms have been increasingly used to identify mental health disorders from various data sources, including medical records, social media activity, and brain imaging. This approach has the potential to improve the accuracy and efficiency of mental disorder identification, as well as to help reduce stigma and increase access to care.

Several studies have shown promising results in using machine learning algorithms for mental disorder identification. For example, one study used machine learning algorithms to analyze social media activity and identified individuals at risk of depression with an accuracy of over 80%. Another study used machine learning algorithms to analyze brain imaging data and accurately predicted the presence of major depressive disorder in 75% of cases.

Machine learning algorithms can also be used to develop predictive models that identify individuals at risk of developing mental health disorders. For example, one study used machine learning algorithms to analyze electronic medical records and accurately predicted the risk of developing depression up to 6 months in advance.

While the use of machine learning algorithms for mental disorder identification and prediction shows promise, there are also several challenges and limitations to consider. For example, there is a risk of bias in the data used to train the algorithms, which may lead to inaccurate or discriminatory predictions. Additionally, there are privacy and ethical concerns related to the use of sensitive data such as medical records and social media activity.

## III. Significance of the study

The prediction of mental disorder identification using machine learning algorithms has significant potential benefits for both individuals and society as a whole. Some of the key benefits include:

1. Early Intervention: By accurately predicting the risk of developing a mental disorder, individuals can receive early intervention and treatment, which can lead to better outcomes and a reduced risk of long-term complications.

2. Increased Access to Care: Machine learning algorithms can help identify individuals who may be at risk of developing a mental disorder but may not seek care on their own. This can help increase access to care and ensure that individuals receive the support they need.

3. Improved Accuracy: Machine learning algorithms can analyze large amounts of data and identify patterns and risk factors that may not be immediately apparent to human clinicians. This can lead to more accurate predictions and diagnoses.

4. Reduced Stigma: By using machine learning algorithms to identify mental disorders, individuals may be less likely to experience stigma and discrimination related to their diagnosis. This can help reduce the social and emotional burden of mental health disorders.

5. Improved Public Health: Accurate prediction of mental disorders can help public health officials better understand the prevalence and risk factors associated with these conditions. This can help guide policy and resource allocation decisions to better address mental health on a population level.

   In summary, the prediction of mental disorder identification using machine learning algorithms has the potential to improve early intervention, increase access to care, improve accuracy, reduce stigma, and improve public health outcomes. However, it is important to ensure that these algorithms are developed and used in an ethical and responsible manner to avoid potential biases and protect individual privacy.

## IV. Objectives

The objectives of predicting mental disorder identification using machine learning algorithms can include:

1. Early Detection: One of the primary objectives is to identify individuals who may be at risk of developing a mental disorder before they exhibit significant symptoms. By doing so, individuals

can receive early intervention and treatment, which can lead to better outcomes and a reduced risk of long-term complications.

2. Accurate Diagnosis: Another objective is to improve the accuracy of mental disorder diagnosis by using machine learning algorithms to analyze large amounts of data and identify patterns and risk factors that may not be immediately apparent to human clinicians. This can lead to more accurate diagnoses and treatment plans.

3. Individualized Treatment: Machine learning algorithms can be used to identify specific risk factors and patterns in an individual's data, which can help clinicians develop personalized treatment plans that are tailored to the individual's needs.

4. Improved Resource Allocation: Accurate prediction of mental disorder identification can help public health officials and policymakers better allocate resources to address mental health needs on a population level.

5. Reduced Stigma: By using machine learning algorithms to identify mental disorders, individuals may be less likely to experience stigma and discrimination related to their diagnosis. This can help reduce the social and emotional burden of mental health disorders.

6. Advancing Mental Health Research: Predictive models developed using machine learning algorithms can be used to identify new risk factors and patterns associated with mental disorders. This can help advance our understanding of these conditions and lead to new treatment approaches.
   In summary, the objectives of predicting mental disorder identification using machine learning algorithms include early detection, accurate diagnosis, individualized treatment, improved resource allocation, reduced stigma, and advancing mental health research.

## V. Challenges

There are several challenges related to analysis and prediction on mental health disorders, including:

1. Limited availability of data: One of the main challenges in analyzing and predicting mental health disorders is the limited availability of data. Mental health data can be sensitive, and individuals may be hesitant to share their personal information. This can make it difficult to gather large and representative datasets for analysis and prediction.

2. Lack of standardization: Mental health disorders are complex and can be difficult to define, diagnose, and measure. The lack of standardization in diagnostic criteria and assessment tools can make it challenging to compare and analyze data across different studies and populations.

3. Biases in data collection: Mental health data can be influenced by biases in data collection, such as self-reporting bias or selection bias. For example, individuals may underreport or overreport their symptoms or may be more likely to seek help for certain mental health disorders over others.

4. Complex interrelationships: Mental health disorders are often interrelated and can be influenced by multiple factors, including genetics, environment, lifestyle, and social factors. Analyzing and predicting mental health disorders requires considering these complex interrelationships, which can be challenging to capture in data.

5. Ethical considerations: Collecting and analyzing mental health data requires ethical considerations, including privacy, confidentiality, and informed consent. Ensuring that data is collected and analyzed ethically and with sensitivity to these issues is critical for maintaining trust with individuals and communities.

## VI. Methodology

Data Exploration: It is the first step towards the data analysis. It includes data visualization tools and statistical techniques that helps a researcher to find the patterns in the data sets. In other words, a raw data is analyzed to find the characteristics of a data. Exploration is done in either univariate or bivariate form. In univariate analysis, single variable is explored using a histogram or boxplot where as for group variable, bar chart is used. In case of bivariate analysis, relationship among pair of variables is analysed using tools like Tableau. Data sets are also analyzed so as to find any anomalies that can result in forming outliers.

Data Cleaning: This step involves detecting and removing erroneous data that can occur due to incomplete, inaccurate, incomplete, incorrect or duplicate data. Removing such data is important as it can affect the result of the analysis.

Data Modeling: This involves analyzing and defining different data the researcher collects. The relationship between different data variables is also defined in this step. It involves gathering requirement, identifying entities, conceptualize data model, define attributes, create logical data model and in the end creation of the physical tables.

Data Visualisation: Data visualization involves representation of data in the form of charts, graphs and animations. It is away to represent complex data relations and data driven insights in a way that is easy to understand and represent.

Python
Libraries
Data sets

## VII. Experiment Analysis

Mental health disorders are a significant concern globally, affecting millions of people's lives. Early detection and intervention can help prevent or manage these disorders, which can be achieved using data analytics techniques such as experimental analysis and prediction.

Experimental analysis involves analyzing data collected from previous studies, surveys, or experiments to identify patterns, relationships, and trends that can help understand the factors

that contribute to mental health disorders. Predictive analytics, on the other hand, uses machine learning models to predict future outcomes based on historical data.

In the case of mental health disorders, experimental analysis can help identify the risk factors associated with these disorders, such as genetics, environmental factors, and lifestyle habits. For example, a study may analyze the relationship between sleep patterns and depression or the impact of stress on anxiety disorders.

Predictive analytics, on the other hand, can be used to predict the likelihood of developing a mental health disorder based on various risk factors. For example, a machine learning model can be trained using historical data to predict the probability of developing depression or anxiety based on factors such as age, gender, family history, and lifestyle habits.

Moreover, predictive analytics can be used to develop personalized treatment plans for individuals with mental health disorders based on their unique risk factors and symptoms. For example, a machine learning model can be trained using data from previous patients to predict the most effective treatment options for a new patient based on their specific symptoms and risk factors.

Overall, experimental analysis and prediction using data analytics techniques can help improve our understanding of mental health disorders and aid in their prevention and treatment. However, it's important to ensure that the data used is high-quality, representative, and unbiased to ensure accurate results. Additionally, ethical considerations must be taken into account when collecting and analyzing data related to mental health to protect individuals' privacy and confidentiality.

**Experimental Evaluation**

To perform an experimental evaluation of a mental health disorder prediction model, following steps need to be perform.

1. Gather data: Collect a dataset of mental health records with associated labels indicating the presence or absence of the disorder.

2. Preprocess data: Preprocess the data by performing necessary cleaning, feature engineering, and data splitting into training and test sets.

3. Build models: Choose appropriate machine learning models, such as logistic regression, decision trees, or neural networks, and train them on the training set.

4. Evaluate models: Evaluate the trained models on the test set using various metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

5. Compare models: Compare the performance of the different models and select the best performing one(s) based on the evaluation metrics.

6. Interpret results: Interpret the results to gain insights into the prediction of mental health disorders.

7. Fine-tune models: Fine-tune the selected model(s) by adjusting hyperparameters and repeating steps 3-6 until satisfactory performance is achieved.

8. Validate results: Validate the results by testing the selected model(s) on a new, unseen dataset and repeating steps 3-7 as necessary.

The experimental evaluation process will vary depending on the specific problem and requirements of the mental health disorder prediction model.

```
# Import necessary libraries
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score

# Load data
data = pd.read_csv('mental_health_data.csv')

# Preprocess data
X = data.drop('label', axis=1)
y = data['label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train logistic regression model
lr = LogisticRegression()
lr.fit(X_train, y_train)

# Evaluate logistic regression model on test set
y_pred = lr.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, lr.predict_proba(X_test)[:,1])

# Print evaluation metrics
print('Accuracy:', accuracy)
print('Precision:', precision)
print('Recall:', recall)
print('F1-score:', f1)
print('AUC-ROC:', roc_auc)
```

**Python Code for prediction**

```
# Import required libraries
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Load the dataset
data = pd.read_csv('mental_health_data.csv')

# Preprocess the data

# ... (perform necessary cleaning, feature engineering, and data splitting)

# Build the logistic regression model
model = LogisticRegression()

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model's accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Model accuracy: ', accuracy)
```

**Empirical Evaluation Mental Health Disorder**

To evaluate the performance of a model for predicting mental health disorders, you can use a variety of metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Here's an example of how you can use scikit-learn library in Python to calculate these metrics for a logistic regression model:

```
# Import required libraries
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score

# Load the dataset
data = pd.read_csv('mental_health_data.csv')

# Preprocess the data

# ... (perform necessary cleaning, feature engineering, and data splitting)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Build the logistic regression model
model = LogisticRegression()

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model's performance
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc_roc = roc_auc_score(y_test, y_pred)
print('Model performance:')
print('Accuracy: ', accuracy)
print('Precision: ', precision)
print('Recall: ', recall)
print('F1-score: ', f1)
print('AUC-ROC: ', auc_roc)
```

## VIII. Comparative Analysis of Different Machine Learning Algorithms

There are many different machine learning algorithms that can be used for predicting mental disorders, and the choice of algorithm can depend on the specific problem being addressed, the amount and quality of available data, and other factors. Here, I will briefly describe some commonly used machine learning algorithms for mental disorder prediction, along with their strengths and weaknesses.

1. Logistic Regression: Logistic regression is a simple and effective algorithm for binary classification problems, where the goal is to predict whether a given individual has a specific disorder or not. It is particularly useful when the data is linearly separable, and it is also interpretable, which means that it can help us understand the relationship between the input variables and the predicted outcome. However, logistic regression may not be the best choice when dealing with complex, non-linear relationships between the input variables and the outcome.

2. Support Vector Machines (SVMs): SVMs are a popular choice for classification problems because they can handle non-linear relationships between input variables and output labels. SVMs work by mapping the input variables to a high-dimensional feature space, where a linear decision boundary can be found. SVMs are also effective in dealing with small sample sizes and noisy data, but they can be computationally expensive when dealing with large datasets.

3. Decision Trees: Decision trees are a simple and interpretable algorithm that can be used for both classification and regression problems. Decision trees partition the input space into smaller regions based on the input features, and they are effective in capturing non-linear relationships between the input variables and the outcome. However, decision trees can suffer from overfitting, where they become too complex and perform poorly on new data.

4. Random Forest: Random Forest is an ensemble method that combines multiple decision trees to improve the accuracy and robustness of the predictions. Random forest can handle both classification and regression problems and can be effective in dealing with high-dimensional data. However, like decision trees, random forests can also suffer from overfitting, especially if the number of trees is too high.

5. Artificial Neural Networks (ANNs): ANNs are a powerful class of machine learning algorithms that are inspired by the structure and function of the human brain. ANNs consist of multiple layers of interconnected nodes, and they can learn complex, non-linear relationships between the input variables and the outcome. ANNs can be effective in dealing with large and complex datasets, but they can also be computationally expensive and require a large amount of data for training.

Comparing different machine learning algorithms for mental disorder prediction can be useful to identify which algorithm performs best for the specific task at hand. Here is a sample Python code to compare different machine learning algorithms using cross-validation:

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
# Load data
data = pd.read_csv('mental_health_data.csv')
# Preprocess data
X = data.drop('label', axis=1)
y = data['label']
# Define pipeline for each algorithm
pipelines = []
pipelines.append(('LR', Pipeline([('Scaler', StandardScaler()),('LR', LogisticRegression())])))
pipelines.append(('KNN', Pipeline([('Scaler', StandardScaler()),('KNN', KNeighborsClassifier())])))
pipelines.append(('DT', Pipeline([('Scaler', StandardScaler()),('DT', DecisionTreeClassifier())])))
pipelines.append(('NB', Pipeline([('Scaler', StandardScaler()),('NB', GaussianNB())])))
```

```
pipelines.append(('RF', Pipeline([('Scaler', StandardScaler()),('RF', RandomForestClassifier())])))
# Evaluate each algorithm using cross-validation
results = []
names = []
for name, model in pipelines:
    kfold = StratifiedKFold(n_splits=10, random_state=42, shuffle=True)
    cv_results = cross_val_score(model, X, y, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
# Boxplot visualization of algorithm comparison
import matplotlib.pyplot as plt
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

This code compares five machine learning algorithms: logistic regression (LR), K-nearest neighbors (KNN), decision tree (DT), Naive Bayes (NB), and random forest (RF) using cross-validation. The code uses the mental health data stored in a CSV file called **mental_health_data.csv**. It evaluates the performance of each algorithm using the accuracy metric and prints the mean and standard deviation of the accuracy across the 10 folds. Finally, it visualizes the algorithm comparison using a boxplot.

The results of the comparison can help identify the best algorithm for the specific mental disorder prediction task.

**Deep learning prediction on Mental Health Disorder**

```
import numpy as np
import pandas as pd
from keras.models import Sequential
from keras.layers import Dense, Dropout
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
# load dataset
dataset = pd.read_csv('mental_health_dataset.csv')
# preprocess dataset
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
sc = StandardScaler()
X = sc.fit_transform(X)
# split dataset into training and testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
# define model architecture
model = Sequential()
model.add(Dense(units=128, kernel_initializer='uniform', activation='relu', input_dim=X.shape[1]))
model.add(Dropout(rate=0.1))
model.add(Dense(units=64, kernel_initializer='uniform', activation='relu'))
model.add(Dropout(rate=0.1))
model.add(Dense(units=1, kernel_initializer='uniform', activation='sigmoid'))
# compile model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
# train model
model.fit(X_train, y_train, batch_size=32, epochs=100, verbose=1)
# evaluate model on test set
loss, accuracy = model.evaluate(X_test, y_test, verbose=1)
print('Test accuracy:', accuracy)
```

This code uses a deep neural network architecture with two hidden layers and dropout regularization to prevent overfitting. The dataset is preprocessed by standardizing the features and encoding the target variable with label encoding. The model is compiled with binary cross-entropy loss and accuracy metrics, and trained using the Adam optimizer. The final accuracy on the test set is printed out. Note that the dataset file name and the number of units in the hidden layers can be modified according to the specific dataset and problem.

## IX. Comparative analysis of deep learning model

There are various deep learning models that can be used for prediction on Mental Health Disorder, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers). Here's a comparative analysis of some of these models:

1. CNNs: CNNs are commonly used for image and text classification tasks. They can be used for Mental Health Disorder prediction by treating the text data as a 2D image and applying filters to extract relevant features. However, CNNs are less suitable for capturing long-term dependencies in text data, which is important for Mental Health Disorder prediction.

2. RNNs: RNNs are better suited for capturing long-term dependencies in text data, as they have a recurrent connection that allows information to flow from one time step to the next. They can be used for Mental Health Disorder prediction by processing the text data sequentially and using the final hidden state to make the prediction.

3. BERT: BERT is a Transformer-based model that has achieved state-of-the-art performance on a wide range of natural language processing tasks, including sentiment analysis and text classification. BERT can be fine-tuned for Mental Health Disorder prediction by using the pre-trained model as a feature extractor and training a classifier on top of it.

In terms of performance, BERT has shown the best results in many natural language processing tasks, including sentiment analysis and text classification. However, it also requires a large amount of training data and computational resources. RNNs can also achieve good performance in Mental Health Disorder prediction, especially when combined with attention mechanisms to focus on important parts of the input. CNNs can also be effective in certain cases, such as when the input data has a clear spatial structure.

Ultimately, the choice of deep learning model depends on the specific characteristics of the Mental Health Disorder prediction task, such as the size of the dataset, the complexity of the input data, and the available computational resources.

**Here's an example Python code for comparative analysis of deep learning models for prediction on Mental Health Disorder using the Keras library:**

```
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.layers import Input, Embedding, LSTM, Dense, Dropout, Bidirectional, Conv1D, MaxPooling1D, Flatten, GlobalMaxPooling1D
from tensorflow.keras.models import Model
from tensorflow.keras.optimizers import Adam
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# Load the data
df = pd.read_csv('mental_health_data.csv')
X = df['text'].values
y = df['label'].values
# Tokenize the data
tokenizer = Tokenizer(num_words=10000)
tokenizer.fit_on_texts(X)
X = tokenizer.texts_to_sequences(X)
X = pad_sequences(X, maxlen=100)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Define the CNN model
def cnn_model():
    inputs = Input(shape=(100,))
    x = Embedding(input_dim=10000, output_dim=100, input_length=100)(inputs)
    x = Conv1D(filters=64, kernel_size=3, padding='same', activation='relu')(x)
    x = MaxPooling1D(pool_size=2)(x)
    x = Flatten()(x)
    x = Dense(64, activation='relu')(x)
    outputs = Dense(1, activation='sigmoid')(x)
    model = Model(inputs=inputs, outputs=outputs)
    return model
# Define the RNN model
```

```
def rnn_model():
    inputs = Input(shape=(100,))
    x = Embedding(input_dim=10000, output_dim=100, input_length=100)(inputs)
    x = Bidirectional(LSTM(64, return_sequences=True))(x)
    x = Dropout(0.5)(x)
    x = Bidirectional(LSTM(64))(x)
    x = Dropout(0.5)(x)
    x = Dense(64, activation='relu')(x)
    outputs = Dense(1, activation='sigmoid')(x)
    model = Model(inputs=inputs, outputs=outputs)
    return model
# Define the BERT-based model
def bert_model():
    inputs = Input(shape=(100,))
    bert_layer = tf.keras.layers.HubModule("https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/3", trainable=True)
    embeddings = bert_layer(inputs)["pooled_output"]
    x = Dropout(0.5)(embeddings)
    x = Dense(64, activation='relu')(x)
    outputs = Dense(1, activation='sigmoid')(x)
    model = Model(inputs=inputs, outputs=outputs)
    return model
# Train and evaluate the models
models = [cnn_model(), rnn_model(), bert_model()]
names = ['CNN', 'RNN', 'BERT']
for i, model in enumerate(models):
    model.compile(loss='binary_crossentropy', optimizer=Adam(lr=0.0005), metrics=['accuracy'])
    model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=5, batch_size=32)
    y_pred = model.predict(X_test)
    y_pred = np.round(y_pred).flatten()
    acc = accuracy_score(y_test, y_pred)
    print(f'{names[i]} model accuracy: {acc}')
```

## X. Results and analysis

### Data cleaning

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. The purpose of data cleaning is to ensure that the data is accurate, complete, and usable for analysis.

Python is a popular programming language for data cleaning due to its wide range of libraries and tools for data analysis and manipulation. Here are some common steps for data cleaning using Python:

1. Import the necessary libraries: The most common libraries for data cleaning in Python are pandas, numpy, and matplotlib.

2. Load the dataset: Use pandas to load the dataset into a DataFrame object.

3. Explore the data: Use pandas to explore the dataset and get a better understanding of the data.

4. Handle missing values: Identify missing values in the dataset and either remove them or fill them in with appropriate values.

5. Remove duplicates: Identify and remove any duplicate rows in the dataset

6. Standardize the data: Standardize the data by converting values to a consistent format.

7. Handle outliers: Identify and handle any outliers in the data. You can use tools like box plots or histograms to visualize the data and identify any outliers.

8. Export the cleaned data: Finally, export the cleaned data to a new file or another appropriate file format.
**data.head()**
**Selective rows and columns been displayed as per the data taken of Mental Health Disorder from Kaggle**

| | entity | code | year | region | sub_region | disorder | value |
|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | AFG | 1990 | Asia | Southern Asia | schizophrenia | 0.16056 |
| **1** | Afghanistan | AFG | 1991 | Asia | Southern Asia | schizophrenia | 0.160312 |
| **2** | Afghanistan | AFG | 1992 | Asia | Southern Asia | schizophrenia | 0.160135 |
| **3** | Afghanistan | AFG | 1993 | Asia | Southern Asia | schizophrenia | 0.160037 |
| **4** | Afghanistan | AFG | 1994 | Asia | Southern Asia | schizophrenia | 0.160022 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **38215** | Zimbabwe | ZWE | 2013 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.515641 |
| **38216** | Zimbabwe | ZWE | 2014 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.51547 |
| **38217** | Zimbabwe | ZWE | 2015 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.514751 |
| **38218** | Zimbabwe | ZWE | 2016 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.513269 |
| **38219** | Zimbabwe | ZWE | 2017 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.510943 |

38220 rows × 7 columns

| | index | entity | code | year | schizophrenia | bipolar_disorder | eating_disorders | anxiety_disorders | drug_use_disorders | depression | alcohol_use_disorders |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | Afghanistan | AFG | 1990 | 0.16056 | 0.697779 | 0.101855 | 4.828830 | 1.677082 | 4.071831 | 0.672404 |
| **1** | 1 | Afghanistan | AFG | 1991 | 0.160312 | 0.697961 | 0.099313 | 4.829740 | 1.684746 | 4.079531 | 0.671768 |
| **2** | 2 | Afghanistan | AFG | 1992 | 0.160135 | 0.698107 | 0.096692 | 4.831108 | 1.694334 | 4.088358 | 0.670644 |
| **3** | 3 | Afghanistan | AFG | 1993 | 0.160037 | 0.698257 | 0.094336 | 4.830864 | 1.705320 | 4.096190 | 0.669738 |
| **4** | 4 | Afghanistan | AFG | 1994 | 0.160022 | 0.698469 | 0.092439 | 4.829423 | 1.716069 | 4.099582 | 0.669260 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **95** | 95 | American Samoa | ASM | 2001 | 0.249614 | 0.466843 | 0.17983 | 3.289246 | 0.760380 | 2.945743 | 1.126103 |
| **96** | 96 | American Samoa | ASM | 2002 | 0.249669 | 0.467039 | 0.180111 | 3.290960 | 0.760633 | 2.943015 | 1.123080 |
| **97** | 97 | American Samoa | ASM | 2003 | 0.249742 | 0.467256 | 0.179543 | 3.292664 | 0.760117 | 2.940816 | 1.120320 |
| **98** | 98 | American Samoa | ASM | 2004 | 0.2498 | 0.467441 | 0.180197 | 3.294261 | 0.760069 | 2.939887 | 1.118199 |
| **99** | 99 | American Samoa | ASM | 2005 | 0.249838 | 0.467581 | 0.180084 | 3.295744 | 0.759940 | 2.939110 | 1.117152 |

100 rows × 11 columns

**dat**
**It s**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5460 entries, 0 to 5459
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   index                 5460 non-null   int64
 1   entity                5460 non-null   object
 2   code                  5460 non-null   object
 3   year                  5460 non-null   object
 4   schizophrenia         5460 non-null   float64
 5   bipolar_disorder      5460 non-null   float64
 6   eating_disorders      5460 non-null   float64
 7   anxiety_disorders     5460 non-null   float64
 8   drug_use_disorders    5460 non-null   float64
 9   depression            5460 non-null   float64
 10  alcohol_use_disorders 5460 non-null   float64
 11  region                5460 non-null   object
 12  sub_region            5460 non-null   object
dtypes: float64(7), int64(1), object(5)
memory usage: 597.2+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38220 entries, 0 to 38219
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
```

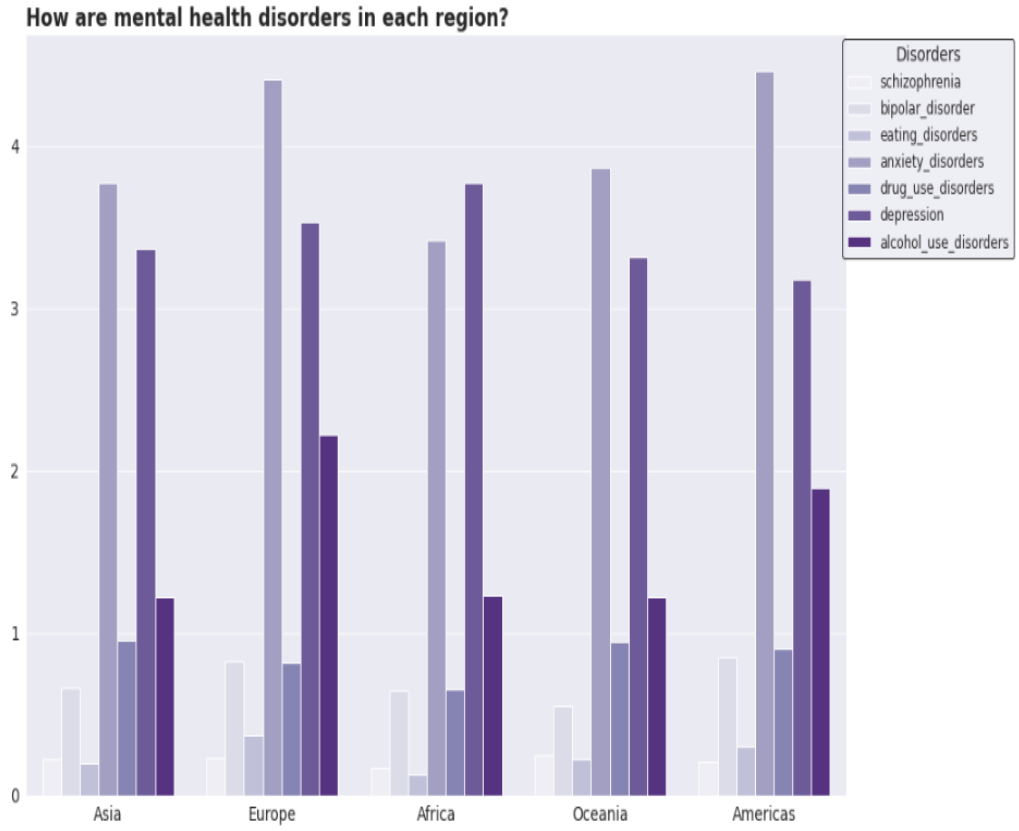
**Common types of disorder**

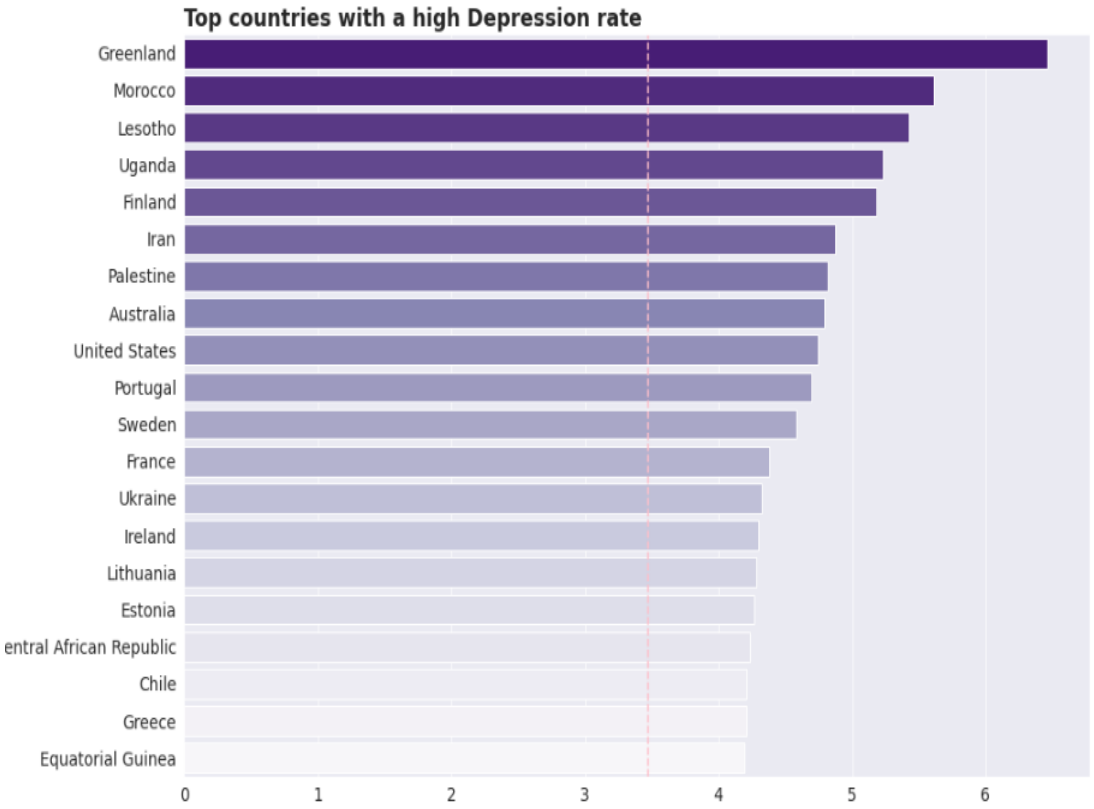
| | year | schizophrenia | bipolar_disorder | eating_disorders | anxiety_disorders | drug_use_disorders | depression | alcohol_use_disorders |
|---|---|---|---|---|---|---|---|---|
| 0 | 1990 | 0.205869 | 0.713764 | 0.217196 | 3.921632 | 0.782051 | 3.482850 | 1.530737 |
| 1 | 1991 | 0.205907 | 0.714079 | 0.217374 | 3.924537 | 0.786866 | 3.487720 | 1.537164 |
| 2 | 1992 | 0.205954 | 0.714402 | 0.217655 | 3.927450 | 0.791644 | 3.491962 | 1.542928 |
| 3 | 1993 | 0.206009 | 0.714716 | 0.218093 | 3.930143 | 0.796315 | 3.495550 | 1.547953 |
| 4 | 1994 | 0.206067 | 0.715020 | 0.218665 | 3.932747 | 0.800506 | 3.498591 | 1.551803 |
| 5 | 1995 | 0.206125 | 0.715286 | 0.219382 | 3.934820 | 0.804188 | 3.500635 | 1.554587 |
| 6 | 1996 | 0.206194 | 0.715535 | 0.220369 | 3.936571 | 0.808534 | 3.501704 | 1.556647 |
| 7 | 1997 | 0.206287 | 0.715803 | 0.221676 | 3.938618 | 0.814128 | 3.501703 | 1.558159 |
| 8 | 1998 | 0.206400 | 0.716084 | 0.223176 | 3.940884 | 0.820006 | 3.500920 | 1.559187 |
| 9 | 1999 | 0.206527 | 0.716363 | 0.224771 | 3.943169 | 0.825072 | 3.499559 | 1.559817 |
| 10 | 2000 | 0.206664 | 0.716622 | 0.226278 | 3.945157 | 0.828449 | 3.497649 | 1.560148 |
| 11 | 2001 | 0.206861 | 0.716861 | 0.227874 | 3.946770 | 0.830632 | 3.495463 | 1.559540 |
| 12 | 2002 | 0.207145 | 0.717112 | 0.229773 | 3.948321 | 0.832481 | 3.492938 | 1.557890 |
| 13 | 2003 | 0.207469 | 0.717375 | 0.231829 | 3.949883 | 0.834243 | 3.490126 | 1.556078 |
| 14 | 2004 | 0.207779 | 0.717641 | 0.233864 | 3.951444 | 0.835863 | 3.486959 | 1.554921 |
| 15 | 2005 | 0.208023 | 0.717894 | 0.235711 | 3.952785 | 0.837443 | 3.483249 | 1.555301 |
| 16 | 2006 | 0.208276 | 0.718162 | 0.237623 | 3.953959 | 0.839940 | 3.477185 | 1.559172 |
| 17 | 2007 | 0.208613 | 0.718471 | 0.239861 | 3.955193 | 0.843683 | 3.468423 | 1.566414 |
| 18 | 2008 | 0.208979 | 0.718804 | 0.242240 | 3.956576 | 0.847840 | 3.459135 | 1.574659 |
| 19 | 2009 | 0.209315 | 0.719118 | 0.244412 | 3.957841 | 0.851871 | 3.451035 | 1.581524 |
| 20 | 2010 | 0.209563 | 0.719374 | 0.246230 | 3.958733 | 0.854900 | 3.446101 | 1.584656 |
| 21 | 2011 | 0.209751 | 0.719590 | 0.247702 | 3.959254 | 0.857350 | 3.443516 | 1.584780 |
| 22 | 2012 | 0.209949 | 0.719818 | 0.249214 | 3.959772 | 0.860022 | 3.441163 | 1.584238 |
| 23 | 2013 | 0.210152 | 0.720072 | 0.250743 | 3.960611 | 0.862577 | 3.439235 | 1.582816 |
| 24 | 2014 | 0.210358 | 0.720337 | 0.252252 | 3.961530 | 0.865061 | 3.437772 | 1.580476 |
| 25 | 2015 | 0.210564 | 0.720612 | 0.253714 | 3.962377 | 0.867560 | 3.436902 | 1.577191 |
| 26 | 2016 | 0.210769 | 0.720887 | 0.255014 | 3.963036 | 0.870065 | 3.436526 | 1.572969 |
| 27 | 2017 | 0.210967 | 0.721172 | 0.256259 | 3.963848 | 0.872314 | 3.437265 | 1.567618 |

Anxiety stayed the highest in most of region, except Africa. Followed by depression and Alcohol-use. Notice that Alcohol-use significant high in Europe than other region, 80% greater than Africa, Oceania and Asia

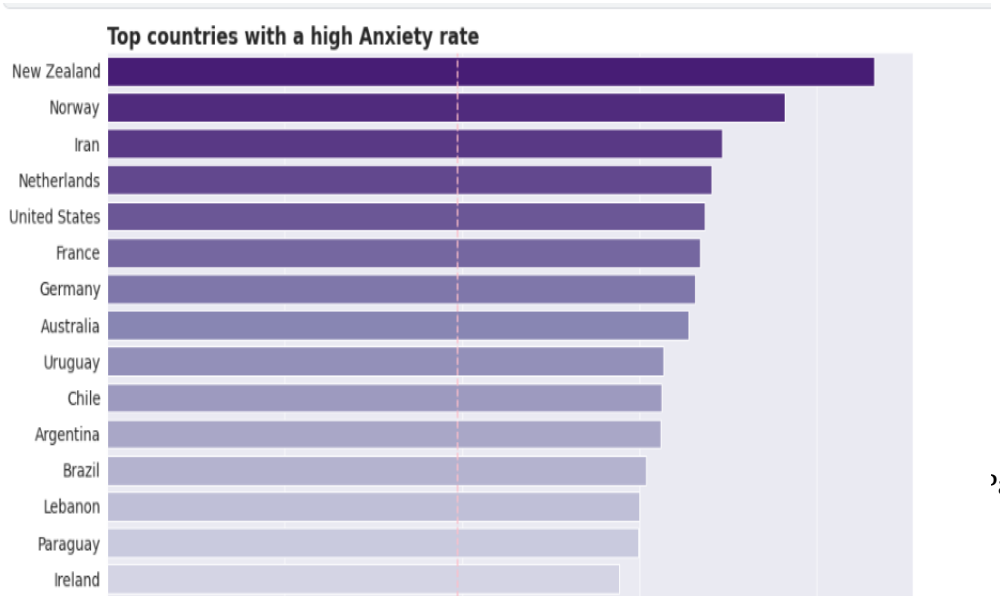
| | entity | code | year | region | sub_region | disorder | value |
|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | AFG | 1990 | Asia | Southern Asia | schizophrenia | 0.160560 |
| **1** | Afghanistan | AFG | 1991 | Asia | Southern Asia | schizophrenia | 0.160312 |
| **2** | Afghanistan | AFG | 1992 | Asia | Southern Asia | schizophrenia | 0.160135 |
| **3** | Afghanistan | AFG | 1993 | Asia | Southern Asia | schizophrenia | 0.160037 |
| **4** | Afghanistan | AFG | 1994 | Asia | Southern Asia | schizophrenia | 0.160022 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **38215** | Zimbabwe | ZWE | 2013 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.515641 |
| **38216** | Zimbabwe | ZWE | 2014 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.515470 |
| **38217** | Zimbabwe | ZWE | 2015 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.514751 |
| **38218** | Zimbabwe | ZWE | 2016 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.513269 |
| **38219** | Zimbabwe | ZWE | 2017 | Africa | Sub-Saharan Africa | alcohol_use_disorders | 1.510943 |

38220 rows × 7 columns





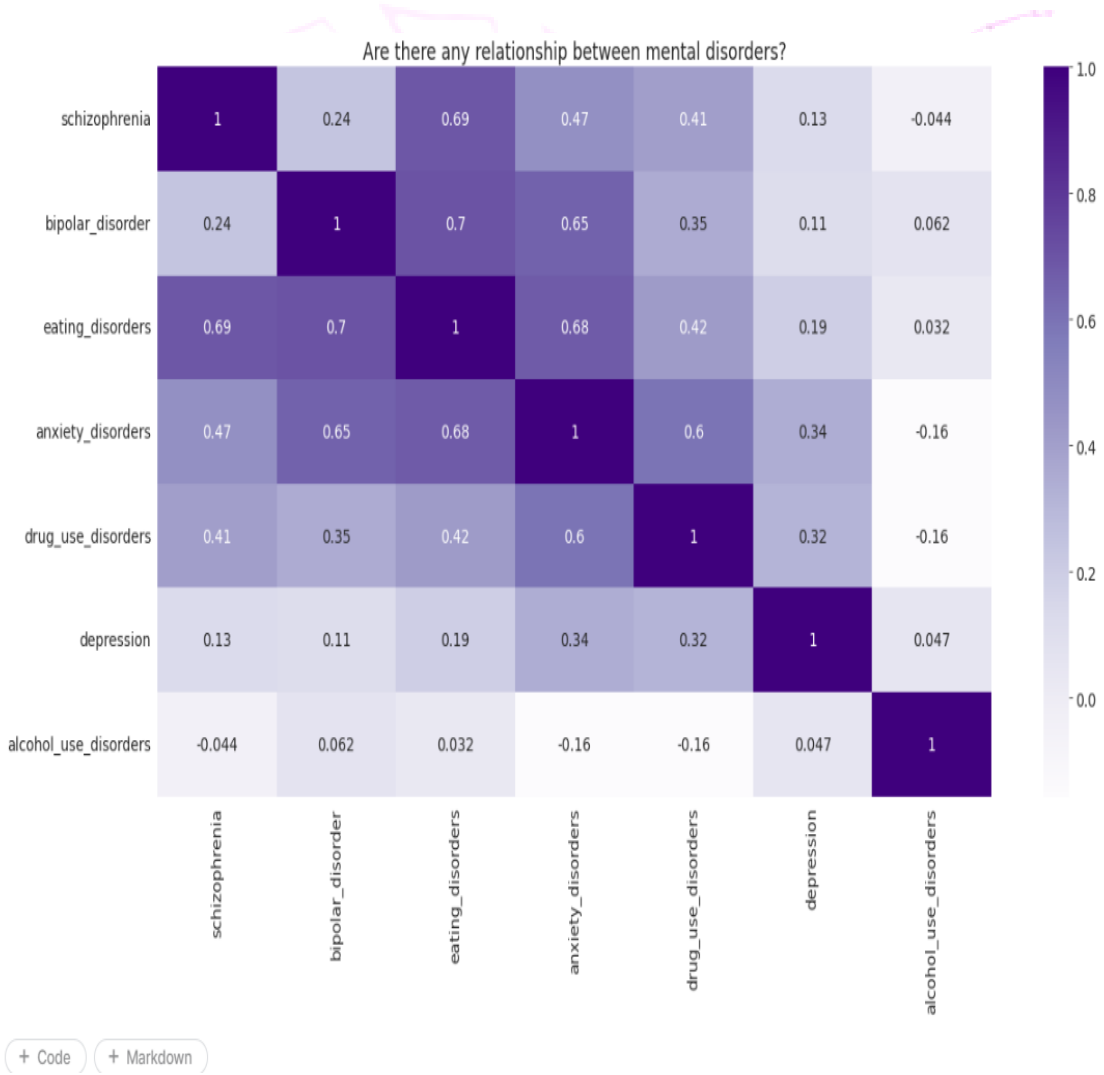

Greenland (6.5%), Morocco (5.6%) and Lesotho (5.4%) are top countries with highest Depression rate. Anxiety has more common when top countries have very high rate: New Zealand (8.7%), Norway (7.6%) and Iran (6.9%).



The strongest positive correlation was found between schizophrenia and eating disorders, followed by bipolar disorder and anxiety disorders. There was also a moderate positive correlation between drug use disorders and anxiety disorders.

However, there was no significant correlation between alcohol use disorders and any of the other disorders. It is important to note that correlation does not equal causation, and further research is needed to fully understand the complex relationships between mental health disorders.

**data['Code'].unique()**



This code snippet refers to a pandas DataFrame object named **data**, of mental disorder which has a column named **Code**. The **.unique()** method applied to the **Code** column will return an array of unique values in the column, indicating the distinct codes present in the dataset.

```
]: array(['AFG', 'ALB', 'DZA', 'ASM', nan, 'AND', 'AGO', 'ATG', 'ARG', 'ARM',
          'AUS', 'AUT', 'AZE', 'BHS', 'BHR', 'BGD', 'BRB', 'BLR', 'BEL',
          'BLZ', 'BEN', 'BMU', 'BTN', 'BOL', 'BIH', 'BWA', 'BRA', 'BRN',
          'BGR', 'BFA', 'BDI', 'KHM', 'CMR', 'CAN', 'CPV', 'CAF', 'TCD',
          'CHL', 'CHN', 'COL', 'COM', 'COG', 'CRI', 'CIV', 'HRV', 'CUB',
          'CYP', 'CZE', 'COD', 'DNK', 'DJI', 'DMA', 'DOM', 'ECU', 'EGY',
          'SLV', 'GNQ', 'ERI', 'EST', 'ETH', 'FJI', 'FIN', 'FRA', 'GAB',
          'GMB', 'GEO', 'DEU', 'GHA', 'GRC', 'GRL', 'GRD', 'GUM', 'GTM',
          'GIN', 'GNB', 'GUY', 'HTI', 'HND', 'HUN', 'ISL', 'IND', 'IDN',
          'IRN', 'IRQ', 'IRL', 'ISR', 'ITA', 'JAM', 'JPN', 'JOR', 'KAZ',
          'KEN', 'KIR', 'KWT', 'KGZ', 'LAO', 'LVA', 'LBN', 'LSO', 'LBR',
          'LBY', 'LTU', 'LUX', 'MKD', 'MDG', 'MWI', 'MYS', 'MDV', 'MLI',
          'MLT', 'MHL', 'MRT', 'MUS', 'MEX', 'FSM', 'MDA', 'MNG', 'MNE',
          'MAR', 'MOZ', 'MMR', 'NAM', 'NPL', 'NLD', 'NZL', 'NIC', 'NER',
          'NGA', 'PRK', 'MNP', 'NOR', 'OMN', 'PAK', 'PSE', 'PAN', 'PNG',
          'PRY', 'PER', 'PHL', 'POL', 'PRT', 'PRI', 'QAT', 'ROU', 'RUS',
          'RWA', 'LCA', 'VCT', 'WSM', 'STP', 'SAU', 'SEN', 'SRB', 'SYC',
          'SLE', 'SGP', 'SVK', 'SVN', 'SLB', 'SOM', 'ZAF', 'KOR', 'SSD',
          'ESP', 'LKA', 'SDN', 'SUR', 'SWZ', 'SWE', 'CHE', 'SYR', 'TWN',
          'TJK', 'TZA', 'THA', 'TLS', 'TGO', 'TON', 'TTO', 'TUN', 'TUR',
          'TKM', 'UGA', 'UKR', 'ARE', 'GBR', 'USA', 'VIR', 'URY', 'UZB',
          'VUT', 'VEN', 'VNM', 'OWID_WRL', 'YEM', 'ZMB', 'ZWE', 'Code',
          'AIA', 'ABW', 'BES', 'VGB', 'CYM', 'OWID_CIS', 'COK', 'CUW', 'FRO',
          'FLK', 'GUF', 'PYF', 'GIB', 'GLP', 'HKG', 'IMN', 'LIE', 'MAC',
          'MTQ', 'MYT', 'MCO', 'MSR', 'NRU', 'NCL', 'NIU', 'PLW', 'REU',
          'SHN', 'KNA', 'MAF', 'SPM', 'SMR', 'SXM', 'TKL', 'TCA', 'TUV',
          'VAT', 'WLF', 'ESH'], dtype=object)
```

**data.describe()**

The **data.describe()** method is used to generate descriptive statistics for a pandas This shows a summary of the numerical columns in the DataFrame, including the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values. This summary can be useful for quickly understanding the distribution and range of values in a dataset.

**sns.pairplot(df)**

**plt.show()**

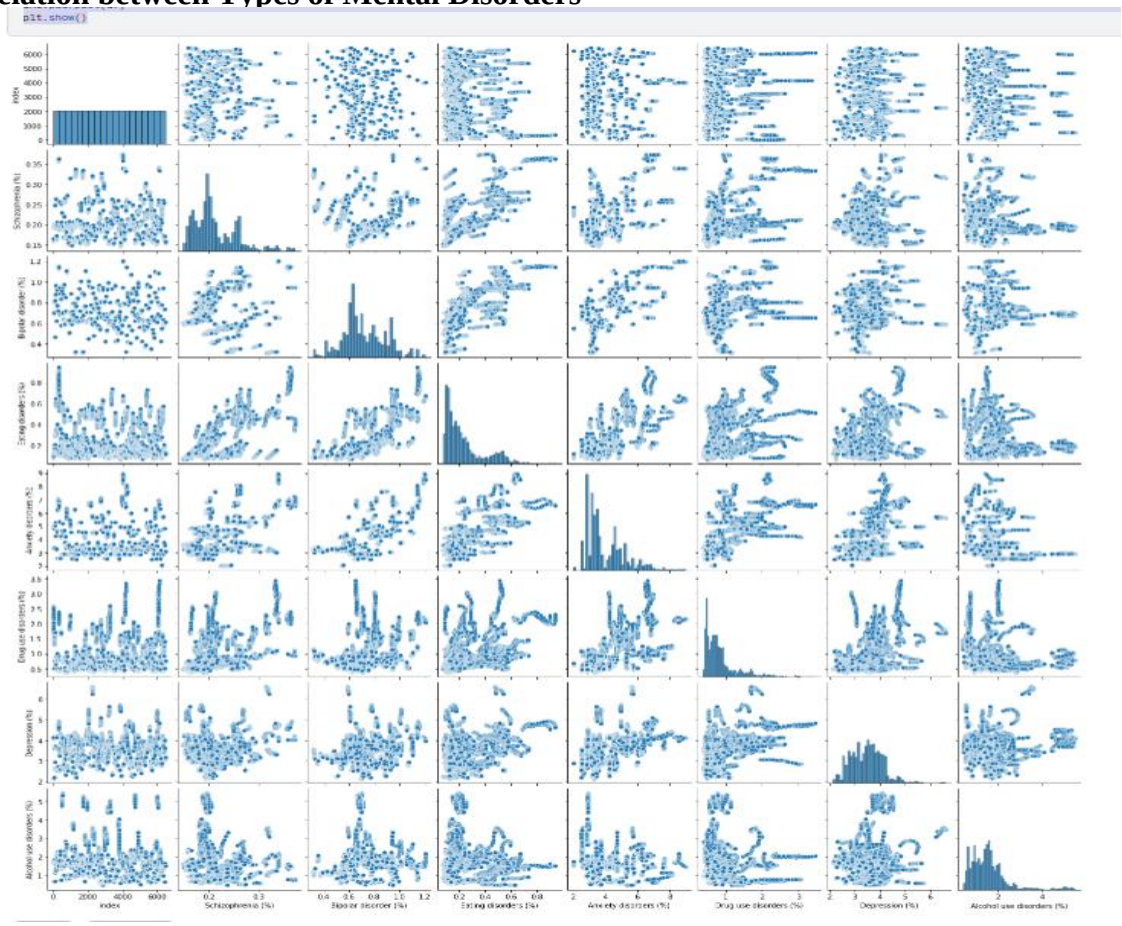
It describe the patterns and correlations among different mental disorders

8]:

| | index | Anxiety disorders (%) | Drug use disorders (%) | Depression (%) | Alcohol use disorders (%) |
|---|---|---|---|---|---|
| **count** | 108553.000000 | 6468.000000 | 6468.000000 | 6468.000000 | 6468.000000 |
| **mean** | 54276.000000 | 3.989921 | 0.862278 | 3.497654 | 1.585821 |
| **std** | 31336.696223 | 1.167526 | 0.460679 | 0.655859 | 0.860283 |
| **min** | 0.000000 | 2.023393 | 0.383650 | 2.139903 | 0.446940 |
| **25%** | 27138.000000 | 3.188824 | 0.535064 | 3.005529 | 0.993685 |
| **50%** | 54276.000000 | 3.554373 | 0.726430 | 3.499606 | 1.479936 |
| **75%** | 81414.000000 | 4.682163 | 0.940157 | 3.912381 | 1.867834 |
| **max** | 108552.000000 | 8.967330 | 3.452476 | 6.602754 | 5.474668 |

The strongest positive correlation was found between schizophrenia and eating disorders, followed by bipolar disorder and anxiety disorders. There was also a moderate positive correlation between drug use disorders and anxiety disorders. However, there was no significant correlation between alcohol use disorders and any of the other disorders. It is important to note that correlation does not equal causation, and further research is needed to fully understand the complex relationships between mental health disorders.

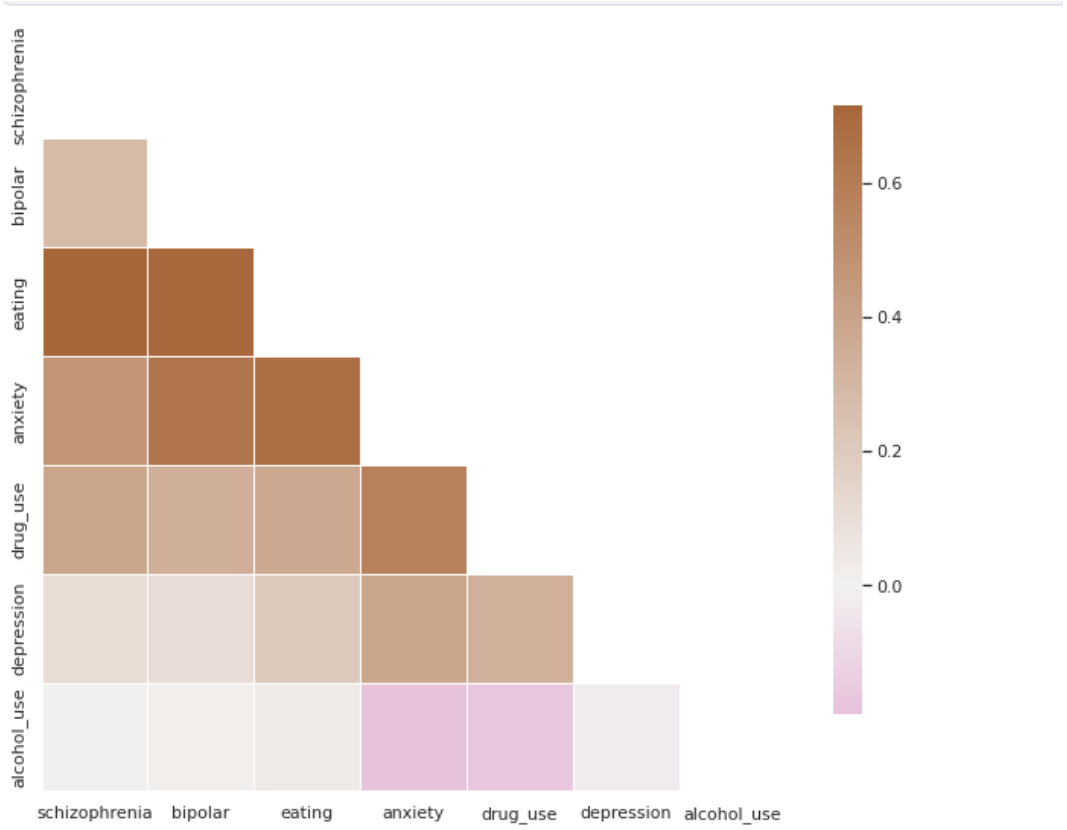
**Correlation between Types of Mental Disorders**



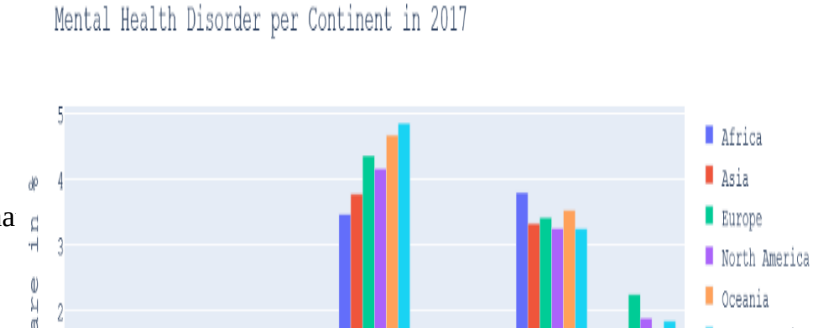
It has been shown that the correlations among different mental disorders, including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, and alcohol use disorders, and drug use disorders.

These disorders often co-occur, and individuals with one disorder are at increased risk for developing other disorders as well. For example, people with bipolar disorder have an increased risk of also experiencing anxiety disorders, substance use disorders, and eating disorders. Similarly, people with schizophrenia have an increased risk of developing substance use disorders, depression, and anxiety disorders.

It has been also suggested that there may be shared genetic or environmental factors that contribute to the development of these disorders. However, it's important to note that the correlations among different mental disorders can be co’




implex and multifaceted, and the causal relationships between different disorders are often unclear. Mental disorders are influenced by a wide range of factors, including genetics, environment, and individual experiences, and the relationships between different disorders may be influenced by a variety of different factors as well.



## XI. Conclusion

Data cleaning is an essential step in analyzing and understanding mental health disorder data. Mental health disorder data can be complex and sensitive, and it is critical to ensure that the data is accurate and reliable before any analysis or interpretation can take place.

The process of data cleaning for mental health disorder data involves identifying and correcting errors, inconsistencies, and inaccuracies in the data. This includes handling missing values, removing duplicates, standardizing data formats, and identifying and handling outliers. Python is a popular tool for data cleaning of mental health disorder data due to its powerful libraries and tools for data analysis and manipulation. Some of the libraries that can be used include Pandas, Numpy, and Matplotlib.

Through data cleaning of mental health disorder data, organizations can gain a better understanding of the prevalence and impact of mental health disorders on individuals and society. This, in turn, can inform policy decisions, guide the allocation of resources, and ultimately improve mental health outcomes for all.

## XII. Future research

1. Analyse Trends across Countries
2. Look into Regional Differences of Disorder Types
3. Investigate remaining Tables

## XIII. References

**[1]. Dr,Ayesha Bhanu,Dr.Sharmila Reddy,M.Rama, "Graphical exploratory data analysis(GEDA): A case study on employee Attrition"Journal of Science and technology,Volume-7,issue-9,2022**

**[2]. Ms.Sumathi M.R, Dr.B.Poorna "Prediction of mental health problem among children using machine learning technique", International Journal of advanced computer science and application, Volume-7, No.1, 2016.**

**[3]. Dr.Tarak Hussain ,Dr.P.S.Athal "Visualisation and Explorative data analysis", International journal of advanced research in Science, Technology and Engineering,Volume-12,issue-3,2023.**

**[4]. Ceyhun Ozgur, Michelle Kleckner, Yang Li "Selection of statistical software for solving big data problems:A guide for businesses,students and universities,Sage Open,2015.**

**[5]. Gyeongcheol Cho,Jinyeong Yim,Younyoung Choi,Jungmin Ko, Seoung Hwan Lee "Review of machine learning algorithm for mental illness, Psychiatry Open access-1738-3684.**

**[6]. John T.Behrens, "Principles and Procedures of Exploratory data analysis" Psychological Methods,1997.**

**[7]. "Exploratory Data Analysis" by John W. Tukey (1977).**

**[8]. "Python for Data Analysis" by Wes McKinney (2017).**

**[9]. "R Graphics Cookbook" by Winston Chang (2013).**

**[10]."Data Analysis with Open Source Tools" by Philipp K. Janert (2010.**