Cairo University

Faculty of Engineering

Systems and Biomedical Department

Spring 2024

Machine Learning(SBE 3230)

Final Project

"Phase 1"

Abdulrahman Emad

Mariam Ahmed Saied

Mourad Magdy

Youssef Ashraf

Under the supervision of

Dr/Inas Yassine

Eng/Merna Bibars

Table of Contents

Definition and Significance

Mental Disorders - Problem Definition

According to WHO:

"mental disorders are known as clinically significant disturbances in an individual's cognition, emotional regulation, or behavior. And It is usually associated with distress or impairment in important areas of functioning."

Diagnosing and treatment of mental illnesses is one of the challenges facing the medical field in this day and age, as mental illnesses can't be diagnosed through lab tests or imaging of the body, which poses a challenge to psychiatrists to identify the root cause of a problem, leading to a problem of common misdiagnosis, "more than a third of patients with severe psychiatric disorders were misdiagnosed (by 39.16%). The commonly misdiagnosed disorder was found to be a schizoaffective disorder (by 75%) followed by major depressive disorder (54.72%), schizophrenia (23.71%), and bipolar disorder (17.78%)."[1]

Thus building classification models that aid with the diagnosis of different disorders will be a step forward to improving the quality of mental health services, and for having more effective treatment planning and intervention.

The dataset provided by the private psychology clinic at Harvard offers a valuable resource for addressing this challenge.[2]

The dataset is categorical, it describes each of the 17 symptoms in ranges, and that's typically how the screening tests used by psychiatrists are made, as mental illnesses mostly lie in terms of spectrums or levels of severity. Understanding the prevalence and severity of symptoms across various conditions can inform more accurate diagnostic processes.

It also serves as a valuable resource for academic research and scientific inquiry into the underlying mechanisms of mental disorders. Researchers can explore correlations between different symptoms, identify risk factors, and uncover novel approaches for intervention and prevention, for individuals classified as "Normal" in the dataset, understanding their utilization of therapy for personal development and enrichment provides insights into preventive strategies and interventions aimed at promoting mental well-being and resilience.

Dataset Description

A Collection of 120 Psychology Patients with 17 Essential Symptoms to Diagnose Mania Bipolar Disorder, Depressive Bipolar Disorder, Major Depressive Disorder, and Normal Individuals. The dataset contains the 17 essential symptoms psychiatrists use to diagnose the described disorders. The behavioral symptoms are considered the levels of patients Sadness, Exhaustness, Euphoric, Sleep disorder, Mood

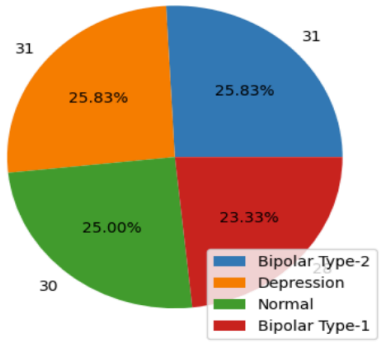
swings, Suicidal thoughts, Anorexia, Anxiety, Try-explaining, Nervous breakdown, Ignore & Move-on, Admitting mistakes, Overthinking, Aggressive response, Optimism, Sexual activity, and Concentration in a Comma Separated Value (CSV) format. The Normal category refers to the individuals using therapy time for specialized counseling, personal development, and life skill enrichment. While such individuals may also have minor mental problems, they differ from those suffering from Major Depressive Disorder and Bipolar Disorder.

Data Preprocessing & EDA

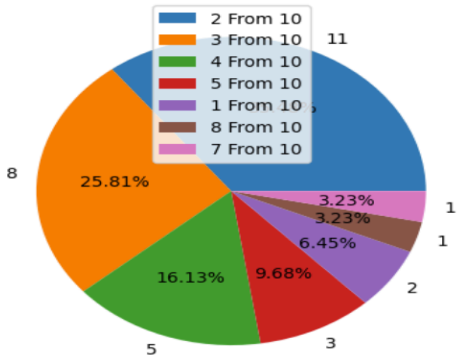
We will drop unnecessary columns such as “Patient Number” to prevent potential noise in the modeling process

Given that the dataset includes categorical features, we need to convert these categorical features into numerical ones. One-hot encoding is a common technique used for this purpose.
After implementing these steps we ensure that the dataset is ready for model training and evaluation.

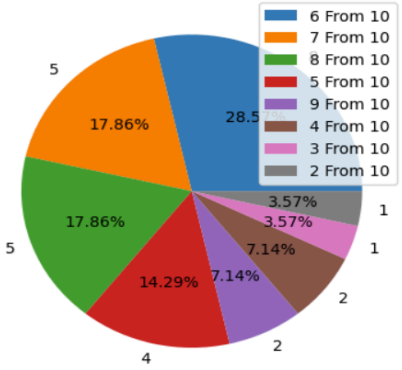
EDA



Symptoms count for each diagnosis



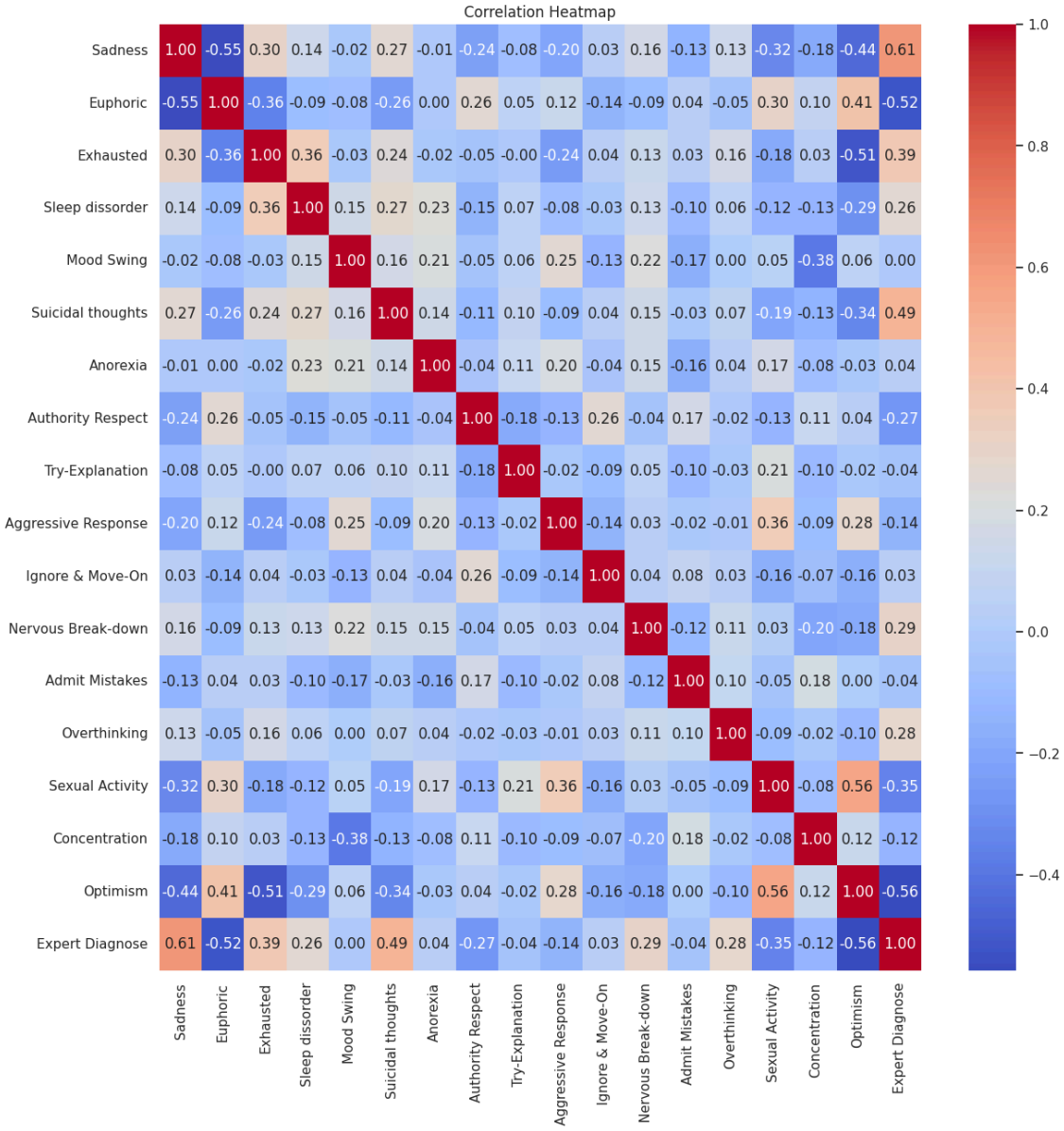
Symptoms count for Bipolar Type-1



Symptoms count for Normal

Feature Extraction

After carefully examining the features heat map, it doesn't seem like any characteristics are being excluded from consideration. All aspects seem to provide useful information, which suggests that all features may be pertinent to the current investigation. Since no obvious candidates for deletion have been found based on the heatmap, it seems wise to keep all features for more review and analysis.



Modeling

We will try several model and choose the best of them, but mainly we will be working with Random Forest and SVM, both are suitable for classification and regression tasks and high dimensional datasets, we will choose the best of them.

Pros of Random Forest:

High Accuracy: Random Forest typically provides high accuracy in classification tasks, often outperforming single decision trees due to its ensemble nature.
Robustness to Overfitting: By averaging multiple decision trees, Random Forest reduces overfitting compared to a single decision tree.

Feature Importance: It provides a measure of feature importance, which can help in feature selection and understanding the data.

Pros of Support Vector Machine:

Effective in High-Dimensional Spaces: SVM performs well even in high-dimensional spaces, making it suitable for tasks like image classification and text classification.

Robust to Overfitting: SVMs are less prone to overfitting, especially in high-dimensional spaces, due to the use of a regularization parameter.

Memory Efficient: SVMs use a subset of training points (support vectors) in the decision function, making them memory efficient, especially when dealing with large datasets.

References

(1) *Ayano G, Demelash S, Yohannes Z, Haile K, Tulu M, Assefa D, Tesfaye A, Haile K, Solomon M, Chaka A, Tsegay L. Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in Ethiopia. Ann Gen Psychiatry.* 2021 Feb 2;20(1):10. doi: 10.1186/s12991-021-00333-7. PMID: 33531016; PMCID: PMC7856725.

(2) *A collection of 120 psychology patients with 17 essential symptoms to diagnose mania bipolar disorder, depressive bipolar disorder, major depressive disorder, and normal individuals*. (2023). [Dataset; Harvard Dataverse]. https://doi.org/10.7910/DVN/0FNET5