Data Collection and Preprocessing Phase

Date	9 July 2024
Team ID	SWTID1720075414
Project Title	Panic Disorder Detection
Maximum Marks	2 Marks

Data Quality Report

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Missing values in the ○ Medical History ○ Psychiatric History ○ Substance Use	Low	Replacing missing values with 'none' `train["Medical History"].fillna("none", inplace=True)` `train["Medical History"].unique()` `array(['Diabetes', 'Asthma', 'none', 'Heart disease'], dtype=object)` `train["Psychiatric History"].fillna("none", inplace=True)` `train["Psychiatric History"].unique()` `array(['Bipolar disorder', 'Anxiety disorder', 'Depressive disorder', 'none'], dtype=object)` `train["Substance Use"].fillna("none", inplace=True)` `train["Substance Use"].unique()` `array(['none', 'Drugs', 'Alcohol'], dtype=object)`
Kaggle Dataset	Panic Disorder Diagnosis column shows a class imbalance	Moderate	Oversampling and resampling SMOTE `print(train["Panic Disorder Diagnosis"].value_counts())` `print(test["Panic Disorder Diagnosis"].value_counts())` `Panic Disorder Diagnosis` `0 95715` `1 4285` `Name: count, dtype: int64` `Panic Disorder Diagnosis` `0 19159` `1 841` `Name: count, dtype: int64` `from imblearn.over_sampling import SMOTE` `smote=SMOTE()` `y_train = train["Panic Disorder Diagnosis"]` `x_train = train.drop(columns=['Participant ID','Panic Disorder Diagnosis'],axis=1)`

			``x_res_train,y_res_train = smote.fit_resample(x_train,y_train)``` ```print(y_train.value_counts()) print(y_res_train.value_counts())``` ```Panic Disorder Diagnosis 0 95715 1 4285 Name: count, dtype: int64 Panic Disorder Diagnosis 0 95715 1 95715 Name: count, dtype: int64```