A project proposal submitted as per the requirements of ST 4052/ DS 4002 – Group 08

Tharindu Fernando – 15522 | Kaveesha Vidushinie – 15572 | Hiruni Kudagama – 15680

Depression Classification

Description of the Problem

Depression is a serious and prevalent mental health disorder which is often characterized by depressed mood, constant feelings of sadness and loss of interest in activities for long periods of time. The gravity of the condition arises from the possibility that it can happen to anyone. Around 280 million people globally suffer from depression, and it is about 50% more prevalent in women compared to men. This justifies the need for an effective tool for depression classification. With the advances of modern technology it is possible to answer this need.

In this project we aim to use machine learning methods to create a system that can accurately identify and classify if a person has depression or not. Various contributing factors such as demographics and lifestyle factors, etc. will be explored along with their influence on the emergence of depression. We aim to fit multiple machine learning models to our data and recognise the best model for depression classification. The findings of our project will benefit healthcare workers and patients alike as they can be used to make informed decisions regarding the management of depression cases.

Description of the Dataset

The " depression" dataset is taken from the Kaggle website and contains 1430 observations. There are 23 variables, out of which the response "depressed" is a categorical variable with two levels (Binary for target class [Zero: No depressed] or [One: depressed]).

Link to the Dataset: https://www.kaggle.com/datasets/diegobabativa/depression/data

Variable Name	Description	Variable Type
Survey_id	Survey Identification number	Nominal-Categorical
Ville_id	Village Identification Number	Nominal-Categorical
sex	Sex of the respondent ; t1-Female, 0- Male	Binary-Categorical
Age	Age of the respondent	Quantitative
Married	Marital status(1-married,0-unmarried)	Binary-Categorical
Number_children	Number of childrens	Ordinal Categorical
education_level	Highest level of education attained by the respondent.	Ordinal Categorical
total_members (in the family)	Total number of family members living in the respondent's household.	Ordinal Categorical
gained_asset	The value of assets gained by the respondent.	Quantitative
durable_asset	The value of durable assets owned by the respondent.	Quantitative

save_asset	The value of savings assets owned by the respondent.	Quantitative
living_expenses	living expenses of the respondent.	Quantitative
other_expenses	Other expenses of the respondent, not including living expenses.	Quantitative
incoming_salary	Indicates whether the respondent has an incoming salary; 1 = Yes, 0 = No.	Binary-Categorical
incoming_own_farm	Indicates whether the respondent has income from their own farm; 1 = Yes, 0 = No.	Binary-Categorical
incoming_business	Indicates whether the respondent has income from a business; 1 = Yes, 0 = No.	Binary-Categorical
incoming_no_business	Indicates whether the respondent has income from non-business sources; 1 = Yes, 0 = No	Binary-Categorical
incoming_agricultural	Income derived from agricultural activities.	Quantitative
farm_expenses	Expenses related to farming activities.	Quantitative
labor_primary	Indicates whether the respondent's primary labor is related to their main income source; 1 = Yes, 0 = No.	Binary-Categorical
lasting_investment	The value of lasting investments made by the respondent.	Quantitative
no_lasting_investmen	Indicates the absence of lasting investments; 1 = No lasting investment, 0 = Presence of lasting investment.	Quantitative
depressed	Zero: No depressed or One: depressed	Binary-Categorical

Comments and/ or Concerns

The dataset's sex distribution (117 male and 1312 female) may skew the analysis results, particularly if there are gender-specific factors influencing depression prevalence or diagnosis. Ensuring gender balance in the dataset or adjusting for gender-related biases is important.In addition to the sex variable, other factors such as marital status (married), incoming salary, and incoming business revenue are also unbalanced. It is essential to examine whether these specific factors influence depression prevalence or diagnosis. Balancing these factors within the dataset or adjusting for related biases is crucial to ensure accurate results.

Moreover, the dataset contains significantly more non-depressed cases (1191) than depressed cases (238), leading to a class imbalance. This imbalance can affect the performance of machine learning models, particularly in accurately predicting the minority class. Techniques such as oversampling, undersampling, or employing appropriate evaluation metrics are necessary to address this issue.Several Machine learning techniques such as XGBoost, KNN, Logistic Regression, MLP, SVM, ADAboost, Gradient Boost will be applied and several evaluation metrics such as accuracy precision F1 score and recall value will be employed for model evaluation purposes.