**DS5110 Project Report**

**Title:** Predictive Analysis for Depression

**Authors (Team 4):**
Rahul Chandak
Dhanooram Nagaraj
Sai Ram Asish Madiraju

**Dataset source:** https://www.kaggle.com/datasets/jeremyteo/predicting-depression-data

**Summary:**

*Problem statement and project goals*

Depression is a widespread problem that affects millions of people all around the world. Detecting and diagnosing it early can be crucial in avoiding the negative impact it can have on people's lives. Many people who are in depression often fail to realize/ accept that they are depressed. This prevents them from taking proper measures and treatments to cope with depression. That's why we're working on a project that aims to predict depression using different algorithms that consider factors like behavioral and psychosocial factors.

*What have we done?*

We have worked on a dataset from Kaggle with entries of 34,364 people who mentioned whether they were depressed or not and a variety of factors that could potentially help us identify any trends which lead to it. Before jumping into this dataset, we contacted our friends and colleagues and undertook a small survey just to identify what variables we will target in addition to the variables which our exploratory data analysis suggests. Further description is covered in the coming sections.

*Description of the dataset:*

The dataset consists of 34,634 rows and 492 columns. The target variable is the 'Depression' columns which determine if someone was depressed or not. The other variables include age, race, weight, employment, income, household size, whether they consumed drugs/ alcohol, health diseases, consumed medicines, etc.

*A non-technical description of the methods:*

- For data tidying, we have replaced missing values in pulse, BMI, and sleep hours with the mean. We removed the rows which had missing values about the user's state (ie whether they were depressed or not depressed).
- In data visualization, we created scatter plots, identified outliers, and __ to find out trends in how depressed people are distributed with age or race, etc.
- To perform correlation analysis, we converted all the categorical variables with numerical variables using an encoder. The target variable is the 'Depression' column which says if a person was depressed or not. We then performed correlation analysis on all the columns with our target variable. Based on the correlation coefficients we shortlisted 30 columns that have a very strong relationship with our target variable.
- Then we used some supervised and unsupervised machine learning algorithms on our filtered dataset and compared which model gave us the highest accuracy.

*Results:*

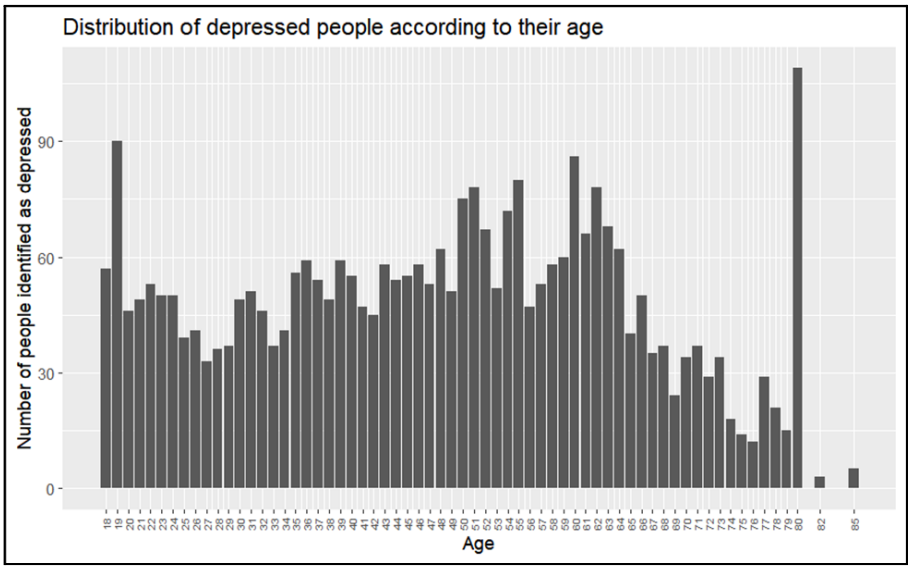
- Single parameters like age or race show no trend or relation with a person being depressed or not. Therefore, we need to take into account a bunch of variables.
- Out of 491 parameters, correlation analysis gives us the top 30 parameters which can be used to determine the outcome.
- On this filtered dataset, the Gradient Boosting classifier model gives us the most accuracy and hence, can be used as a reliable way to make predictions.

**Methods:**

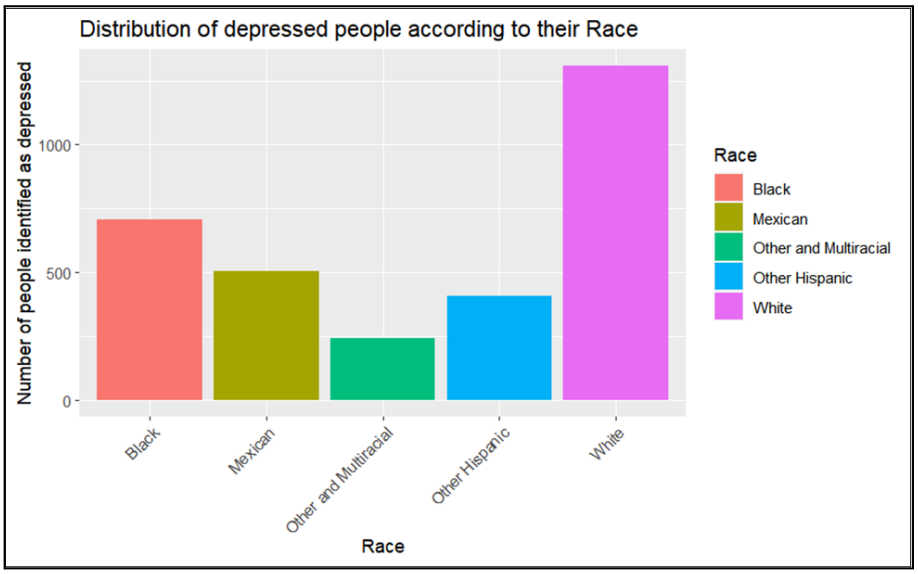
The project methods are attached in a separate pdf.

**Results:**

*Initial observations:*




The above plot represents how the number of depressed people is distributed based on their ages. There is no specific trend observed to target people of an age category. The maximum count is observed for people aged 80 followed by 19.




The above plot represents how the number of depressed people is distributed based on the race they belong to. Most depressed people are observed among whites followed by blacks. However, this alone is not sufficient to conclude that Whites are more prone to being depressed and hence make any predictions.

*Carrying out correlation analysis to shortlist the top 30 factors which are closely related to depressed people*

['Cant Work', 'Memory Problems', 'Limited Work', 'Trouble Sleeping History', 'Health Problem Back Or Neck', 'Out Of Work', 'Walking Equipment', 'Prescriptions Count', 'Healthcare Equipment', 'Health Problem Arthritis', 'Current Cigarettes Per Day', 'Arthritis', 'Health Problem Blood Pressure', 'Rx Gabapentin', 'Bronchitis', 'Health Problem Diabetes', 'Rx Alprazolam', 'Rx Clonazepam', 'Bronchitis Currently', 'Moderate Recreation', 'Health Problem Breathing', 'Health Problem Vision', 'Asthma Currently', 'Health Problem Bone Or Joint', 'Ever Overweight', 'Health Problem Weight', 'Vigorous Recreation', 'Arthritis Type', 'Heart Attack Relative']

*Filtering datasets on these columns and using prediction models*

We fitted this training dataset into 4 different types of models. The details are in the 'Project Methods' pdf. The results are mentioned in the below table:

| **Sr. No.** | **Model Name** | **Accuracy** |
|---|---|---|
| 1 | Support Vector Machine | 91% |
| 2 | Random Forest Classifier | 91.64% |
| **3** | **Gradient Boosting Classifier** | **91.83%** |
| 4 | Naive Bayes | 80.72% |
| 5 | Logistic Regression | 91.74% |

**Discussions:**

Our first approach was to ask around common people (like our colleagues and friends) about what can be potential reasons for a person to be depressed. We shortlisted some parameters which we wanted to analyze in detail:

['Gender', 'Age', 'Race','Marital Status','Pregnant','Household Income','Ever Overweight','Pulse','Sleep Hours','BMI']

In our initial rounds of visualization, we started exploring the dataset based on the following questions:

1. Are people of a particular age group more likely to be victims of depression?
2. Is there a trend in which depressed people belong to a particular race more? Are people of a particular race more prone to being depressed?

Looking at the first two plots from the ***Results*** section, we cannot determine a proper 'age category' which can tell people of what age group is more prone to being depressed. Most depressed people are found to be of ages 19 or 80 but this can be highly influenced because of other parameters. Similarly, from the second plot, we can see that most depressed people belong are 'Whites' compared to the other mentioned races. But we should also keep in mind that the overall population of whites can also be more and hence we cannot generalize this notion that Whites are more prone to being depressed. Therefore, trying to predict depression in people based on 1 or 2 parameters alone cannot give us accurate results.

Therefore, we used correlation analysis, i.e. calculated correlation coefficients and the p values to find out the variables which have the most robust relationship with the 'Depression' column. Both resulted in 30 variables mentioned in the ***Results*** section. Combining those 30 variables with our 10 variables, and then using predicting models helped us conclude that the Gradient Boosting Classifier model can give us the maximum accuracy in predicting depression among an individual based on these 40 variables.

*Who can benefit from our project?*

People of all age groups - from juniors to seniors, students to professionals, everyone can benefit from this project. It can provide them with early warning signs, which can help them seek timely help and treatment.

This project can also help healthcare professionals who can identify patients who are at risk of developing depression and offer them preventive care.

Researchers studying depression can use the data generated by a depression prediction project to identify risk factors and develop new interventions for prevention and treatment.

*Future plans:*

- Use a more diverse dataset. This will help to improve the model's generalizability and reduce bias.
- Involve mental health experts. This helps ensure that the model's outputs are clinically relevant and actionable.
- Incorporate more models and fine-tune existing models to get more details from the dataset and improve accuracy.

**Statement of Contributions:**

**Rahul Chandak:** I was responsible for data tidying and carrying out data visualizations to find some specific trends in the lifestyles of depressed people. Data tidying involved removing NA values and replacing abnormal values (eg. pulse and BMI had 0 values which do not make sense) with the mean of all the values. From the summary statistics and visualizations, I found outliers in the plots which also helped in data cleaning.

**Dhanooram Nagaraj:** I carried out the exploratory data analysis, where I calculated the correlation coefficients and p values of all the parameters with a person being depressed or not. Some variables were categorical which had to be converted to numerical. For this purpose I used labelEncoder. The correlation coefficients helped to shortlist 30 variables that had a strong relationship with our target variable.

**Sai Ram Asish Madiraju:** I was responsible for carrying out the prediction analysis using machine learning models. After feature selection, I trained the filtered dataset on 5 supervised machine-learning models. I observed the confusion matrices, ROC curves, and accuracy of each model and concluded that of all the models, the Gradient Boosting classifier gives the most accurate results (91.83%).

**References:**


1. https://www.kaggle.com/datasets/jeremyteo/predicting-depression-data
2. https://www.nature.com/articles/s41598-021-81368-4
3. https://ibcces.org/blog/2019/05/01/impact-anxiety-depression-student-progress/
4. https://www.publichealth.columbia.edu/news/nearly-one-ten-americans-reports-having-depression#:~:text=Major%20depression%20is%20the%20most,to%207.3%20percent%20in%202015.
5. https://www.ajpmonline.org/article/S0749-3797(22)00333-6/fulltext


**Appendix:**

The entire code is available on the below Github link:

https://github.com/rahulschandak/Predictive-Analytics-for-Depression