greatlearning
Learning for Life

PROJECT SUMMARY:

Batch details	PG-DSE June2021
Team members	G Nikhil Raj Divya P Gaurav Prasad Gond T Vijaya Venkata Krishna
Domain of Project	Health Care
Proposed project title	Study of Factors Leading to Depression
Group Number	1
Team Leader	G Nikhil Raj
Mentor Name	Mr.Ankush Bansal

Date:02-12-2021

Signature of the Mentor

Signature of the Team Leader

Table Of Contents

Industry Review:

Depression is a common illness worldwide, with an estimated 3.8% of the population affected, including 5.0% among adults and 5.7% among adults older than 60 years. Approximately 280 million people in the word have depression. Depression is different from usual mood fluctuations and short-lived emotional responses to challenges in everyday life. Especially when recurrent and with moderate or severe intensity, depression may become a serious health condition. It can cause the affected person to suffer greatly and function poorly at work, at school and in the family. At its worst, depression can lead to suicide. Over 700 000 people die due to suicide every year. Suicide is the fourth leading cause of death in 15-29-year-olds.

Although there are known, effective treatments for mental disorders, more than 75% of people in low- and middle-income countries receive no treatment. Barriers to effective care include a lack of resources, lack of trained health-care providers and social stigma associated with mental disorders. In countries of all income levels, people who experience depression are often not correctly diagnosed, and others who do not have the disorder are too often misdiagnosed and prescribed antidepressants.

The National Survey on Drug Use and Health (NSDUH) provides up-to-date information on tobacco, alcohol, and drug use, mental health, and other health-related issues in the United States. Information from NSDUH is used to support prevention and treatment programs, monitor substance use trends, estimate the need for treatment and inform public health policy.

- **Problem Statements:**

To study the impact of factors leading to depression and predict whether the respondent is prone to depression or not based on the factors like using cocaine, cigarettes, marijuana, crack, income range, education, and many other factors and to build a suitable model so that we can identify the depressed respondents and improve their mental health.

- **Project Outcome:**

This data is a comprehensive survey on drug use and demographics in the United States. The aim of this project is to construct an efficient risk prediction system to detect the possible depressed respondents from the survey which was held by National Survey of Drug Use and Health (2015-2019).

Data Description and Preprocessing

Column name	Description	Type	Sample values
Cigever	Ever used Cigarette	Object	0 - No 1 - Yes
Alcever	Ever Used Alcohol	Object	0 - No 1 - Yes
Crkever	Ever Used Crack	Object	0 - No 1 - Yes
Cocever	Ever Used Cocaine	Object	0 - No 1 - Yes
Herever	Ever Used Heroine	Object	0 - No 1 - Yes
Mjever	Ever used Marijuana	Object	0 - No 1 - Yes
Methanever	Ever Used Methamphetamine	Object	0 - No 1 - Yes
Lsd	Ever Used LSD	Object	0 - No 1 - Yes
Pcp	Ever Used PCP	Object	0 - No 1 - Yes
Peyotte	Ever Used Peyotte	Object	0 - No 1 - Yes
Mesc	Ever used Mesc	Object	0 - No 1 - Yes
Psilcy	Ever Used Psilcy	Object	0 - No 1 - Yes
Ecstmolly	Ever Used Ecstmolly	Object	0 - No 1 - Yes

Ketminesk	Ever used Ketamine or Super K	Object	0 - No 1 - Yes
Dmtamfxy	Ever used DMT (dimenthyltryptamine), AMT (alpha methyltryptamine)	Object	0 - No 1 - Yes
Salviadiv	Ever used Salvia divinorum	Object	0 - No 1 - Yes
Hallucoth	Ever used other Hallucinogen	Object	0 - No 1 - Yes
Inhalever	Ever used inhalants (Amlnit, Cleflu, Gas, Glue, Ether, Solvent, Lgas, Nitoxid, Feltmarkr, Spaint, Airduster, Othaeros or Inhaloth)	Object	0 - No 1 - Yes
Catag3	Age group	Object	1 = 12 – 17 years old 2 = 18 – 25 years old 3 = 26 – 34 years old 4 = 35 – 49 years old 5 = 50 or older
Health	Health Condition	Object	1 = Excellent 2 = Very good 3 = Good 4 = Fair 5 = Poor
Irwrkstat	Work Status	Object	1 = Employed Fulltime 2 = Part-time Employed 3 = Unemployed 4 = Others

Ireduhighst2	Highest Completed Education	float	1 = 5^{th} grade 2 = 6^{th} grade 3 = 7^{th} grade 4 = 8^{th} grade 5 = 9^{th} grade 6 = 10^{th} grade 7 = 11^{th} or 12^{th} grade 8 = High school diploma 9 = Some college credit 10 = Associates degree 11 = College graduate or higher
Newrace2	Race/Ethnicity	Object	1 = Non-Hispanic White 2 = Non-Hispanic Black 3 = Non-Hispanic native American 4 = Non-Hispanic native HI/other Pac Isl 5 = Non-Hispanic Asian 6 = Non-Hispanic more than one race 7 = Hispanic
Irsex	Gender	Object	0 = Female 1 = Male

Irpinc3	Income range	float	1 = < $10000 2 = $10000 - $19999 3 = $20000-$29999 4 = $30000-$39999 5 = $40000-$49999 6 = $50000-$74999 7 = $75000 or more
Irki17_2	Number of kids less than 18y/o	Object	1 = No children under 18 2 = one children under 18 3 = Two children under 18 4 = Three or more children under 18
irmjfy	Number of days used marijuana in past year	float	Range 0-365
iralcfy	Number of days used alcohol in past year	float	Range 0-365
ircocfy	Number of days used cocaine in past year	float	Range 0-365
irherf	Number of days used Heroine in past year	float	Range 0-365
irhallucyfq	Number of days used hallucinogen in past year	float	Range 0-365
irinhalyfq	Number of days used inhalants in past year	float	Range 0 -365
irmethamyfq	Number of days used methamphetamine in past year	float	Range 0 -365
booked	Ever arrested or booked for breaking the law	Object	0-No 1-Yes
probation	On probation at any time in past year	Object	0-No 1-Yes
drvinalco	Drove under the influence of alcohol past 12 months	Object	0-No 1-Yes

drvinamarj	Drove under influence of marijuana	Object	0 -No 1-Yes
drivincocn	Drove under influence of Cocaine	Object	0-No 1-Yes
drvinhern	Drove under influence of Heroine	Object	0 – No 1-YES
drvinhall	Drove under influence of hallucinogen	Object	0-No 1-Yes
drivininhl	Drove under influence of inhalants	Object	0-No 1-Yes
drvinmeth	Drove under the influence of inhalants	Object	0-No 1-Yes
mhsuipln	Made plans to kill self in past year	Object	0-No 1-Yes
mhsuitry	Attempted to kill self in past year	Object	0-No 1-Yes
Wrkdhrswk2	Number of hours worked in past week	float	Range 0-60
Irmarit	Marital Status	Object	1-Married 2-Widowed 3-Divorced/Separated 4-Never married
Coutyp4	Metro/Non-metro	Object	1-Large metro 2-Small metro 3-Non metro
Irhhsiz2	Number of people in household	Object	0-No 1-Yes
Cig30use	Number of days smoked cigarettes in past month	int	Range 1-30
addprev	Have you ever in your life had a period lasting several days or longer when most of the day you felt sad, empty or depressed	Object	0-No 1-Yes

Data Preparation

Data preprocessing is a crucial step that helps enhance the quality of data to promotethe extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models

Data Preparation is the process of collecting, cleaning, and consolidating data intoone file or data table, primarily for use in analysis.

- Acquire the dataset:

  https://www.kaggle.com/bgallamoza/national-survey-of-drug-use-and-health-20152019?select=NSDUH_2015-2019.csv
- Import the dataset.
- Exploratory data analysis
- Identifying and handling the missing values/outliers
- Encoding the categorical data
- Data transformations and scaling using: -'yeo-johnson' from PowerTransformer in the sklearn library
- Statistical adjustments
- Splitting the dataset
- Feature scaling

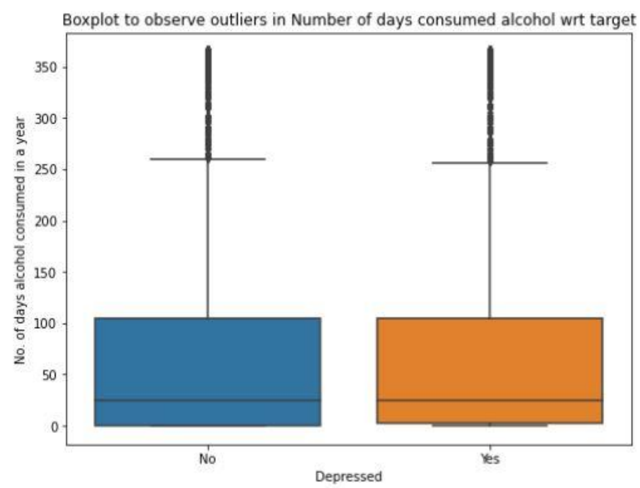
The dataset has 2812 columns and 282768 rows, identifying the target column and the independent columns which are helpful in predicting the target and storing the columns in a single data frame/csv file for use in analysis.

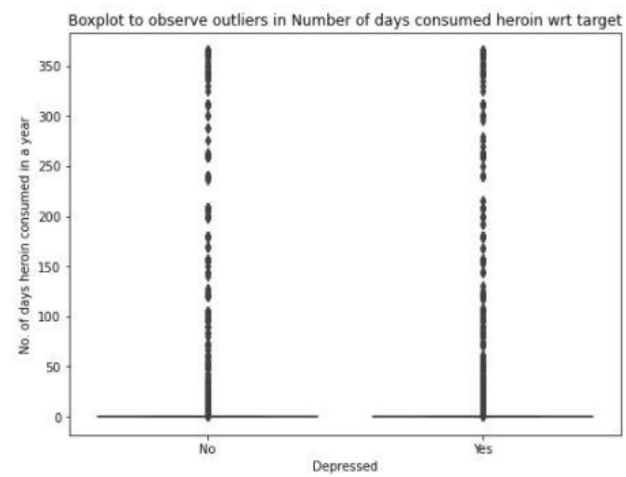
Preprocessing:

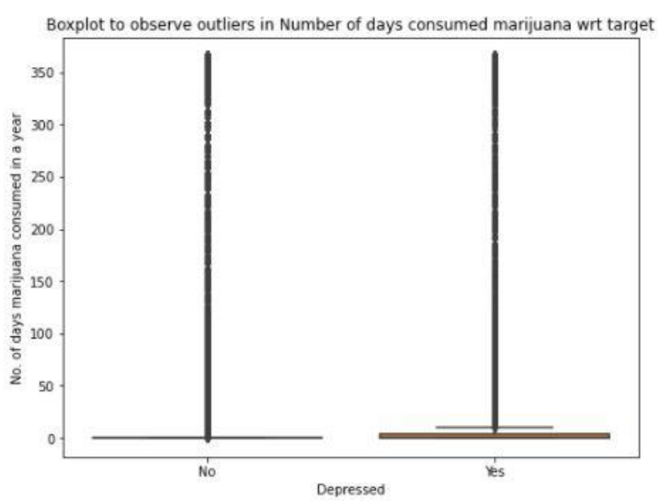
- Null values are observed in all the columns.
- Dropping the null values in the target, as they cannot be predicted.
- After dropping the null values in the target, null values are present in all the remaining columns are also rectified.
- Columns with null value percentage above 25% are dropped.
- For the remaining columns, null values are imputed using KNN imputation.

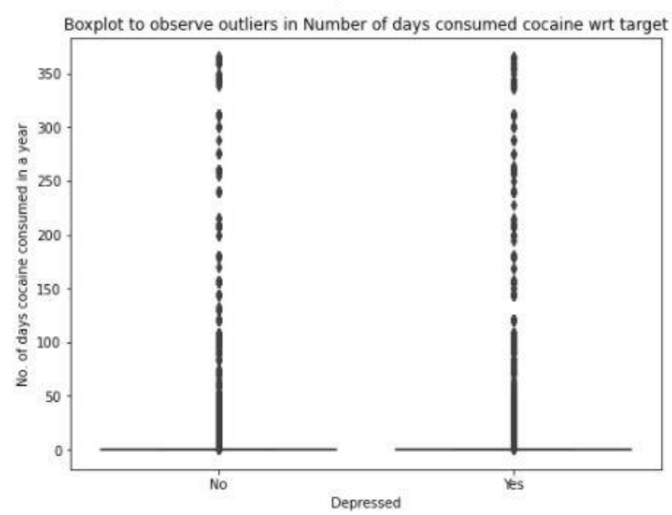
Outlier Treatment:

Outliers are present in the continuous columns, the data is drug related, the outliers are to be considered



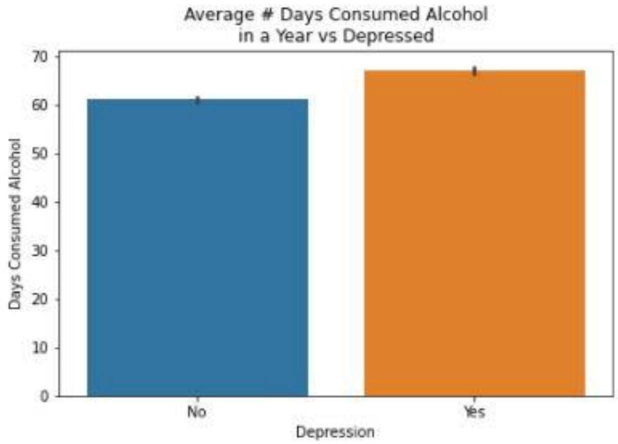


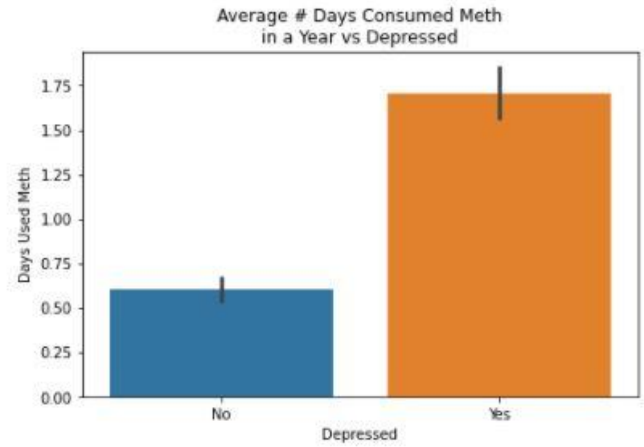




From the above graph we can observe that the target variable, cannot be distinguished in the outliers, this similar for all the continuous variables. Hence the outliers are to be considered in the model building for the analysis.

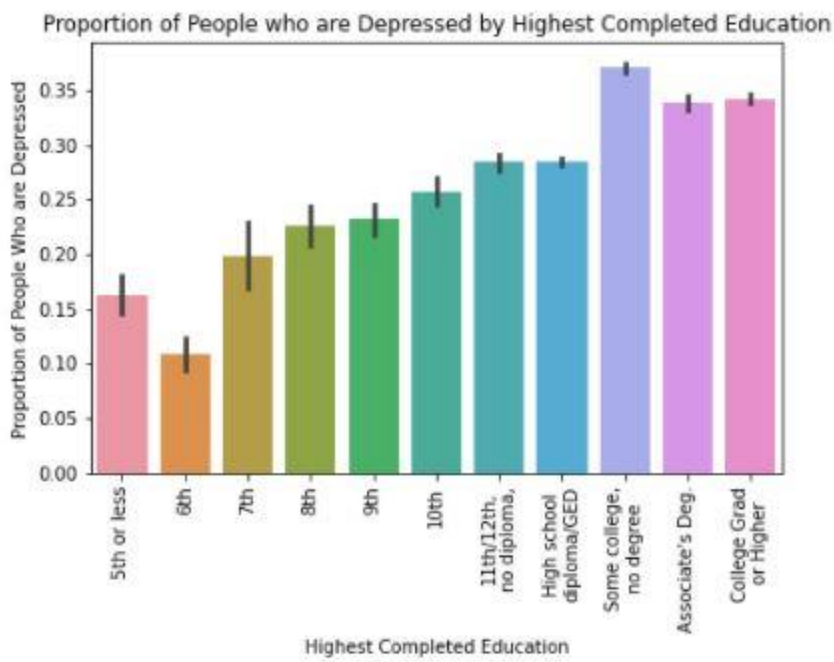
Exploratory Data Analysis & Insights:



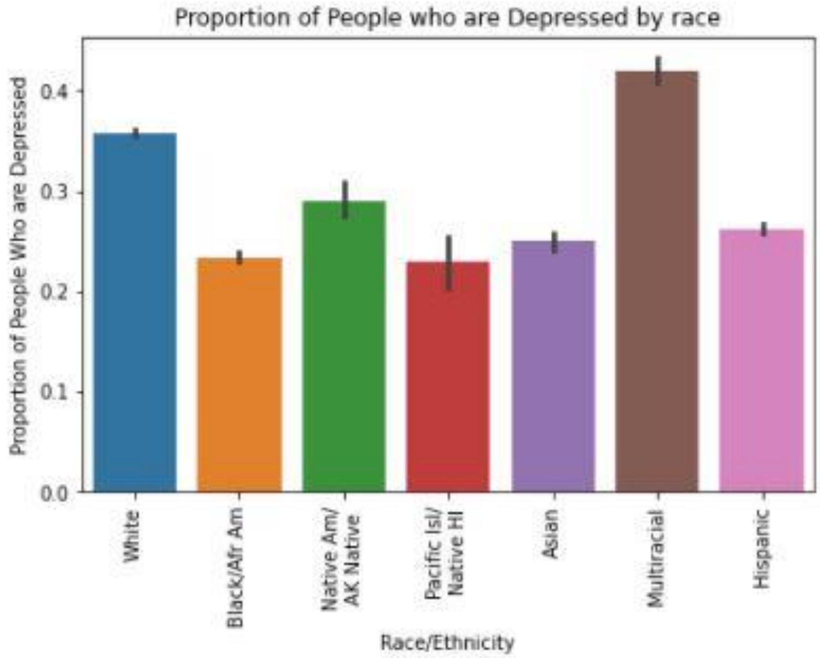


The average consumption of alcohol in a year is similar in for people who are depressed and not depressed.

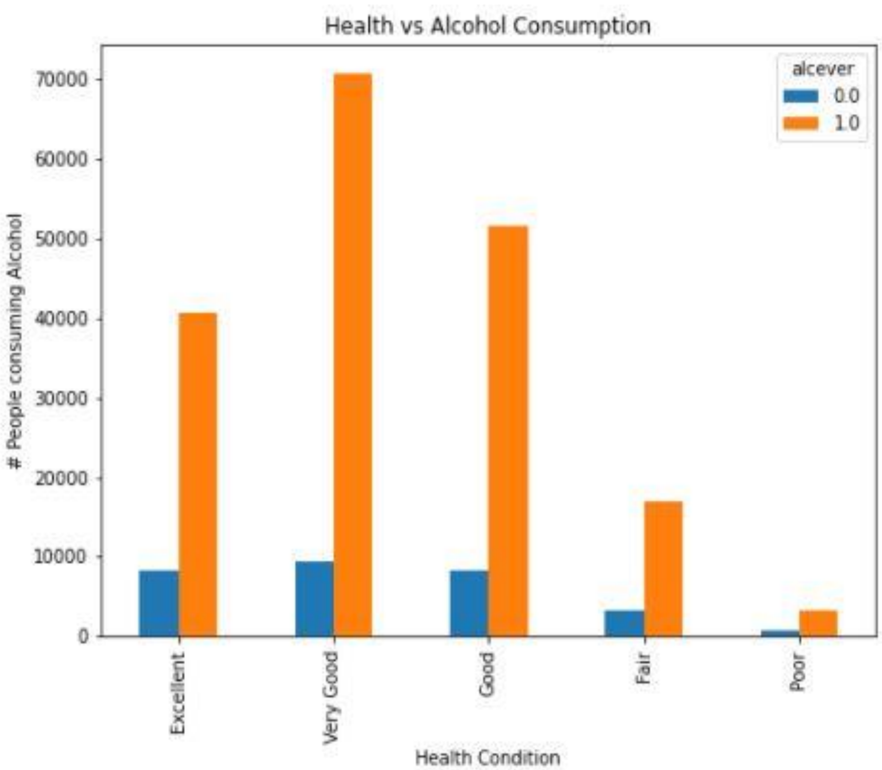
The average consumption of meth in a year is high for people who are depressed and not



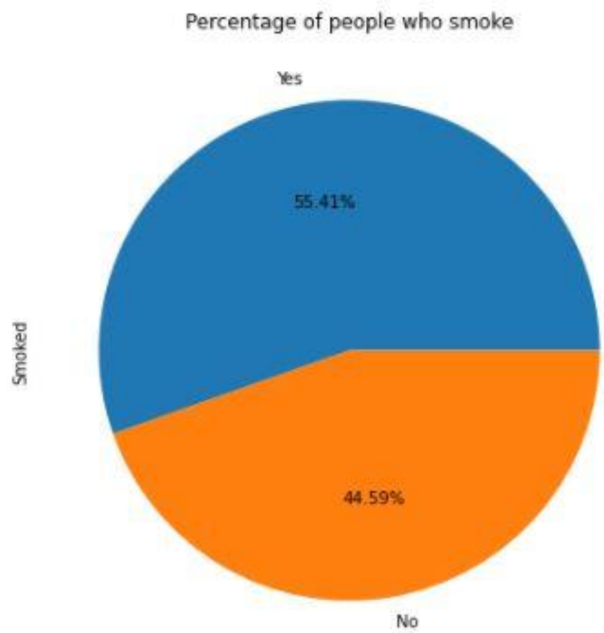
From the above graph we can observe that people who are more educated are facing with mental health/depression problems.

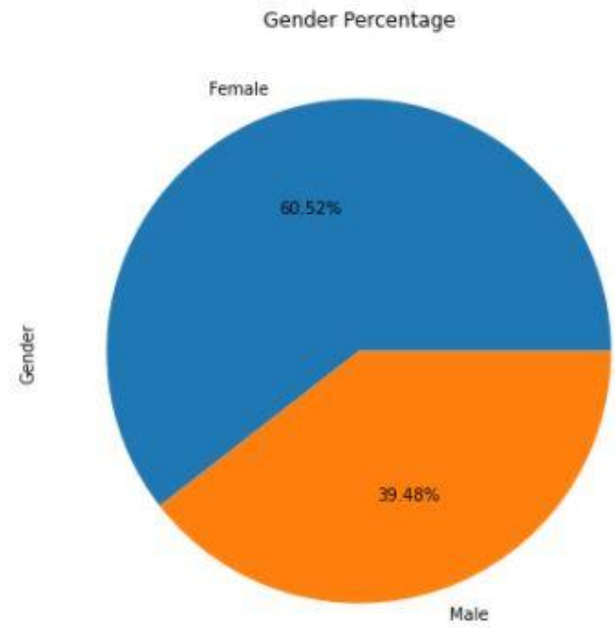


Maximum proportion of people from multiracial are depressed, followed by white, native American, Hispanic, Asian, Black, Pacific Islanders



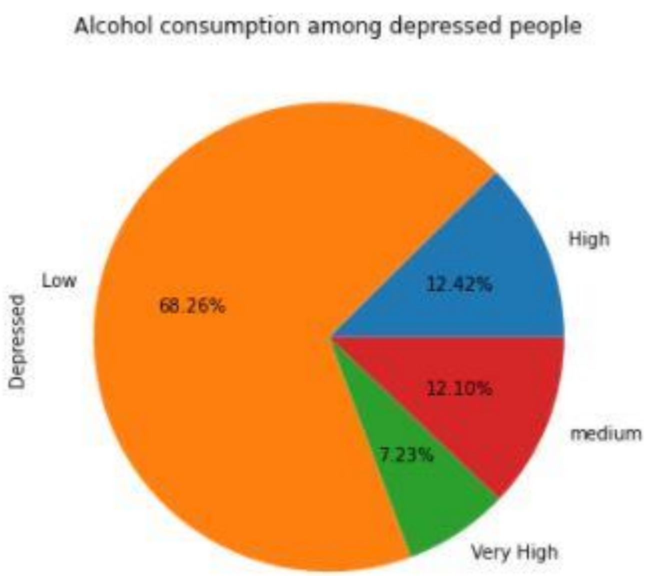
People with very good health condition consume more alcohol, followed by Excellent, good, fair, and poor.

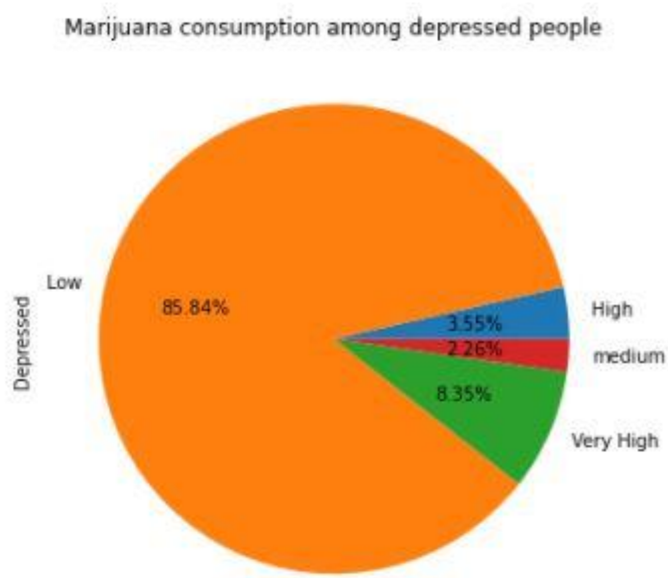




More number of people smoke cigarettes and are prone to depression when compared to people who don’t smoke.

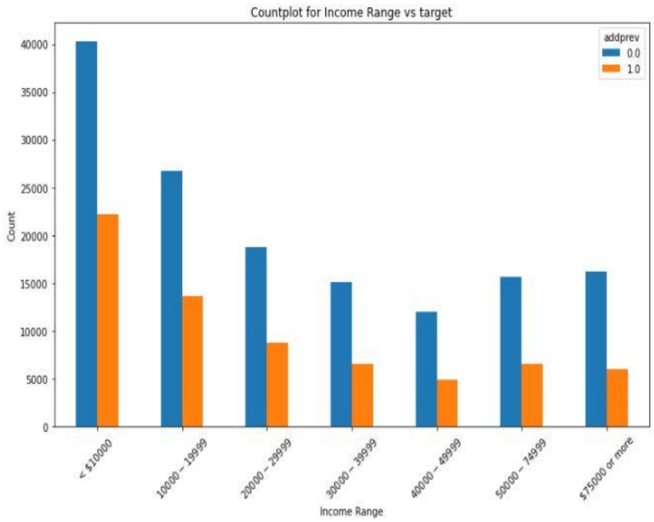
The percentage of females is higher when compared to males who are more prone to depression.

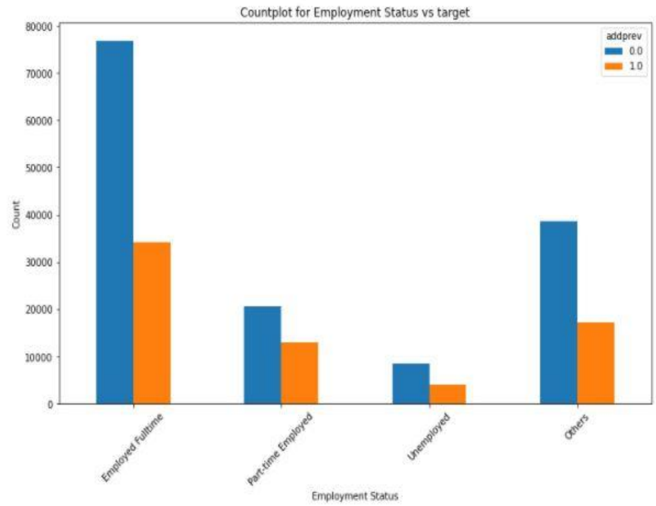




Among depressed people, 12.10% people consume alcohol more than 6 months in a year, 68.26% people consume alcohol less than 2 months in a year

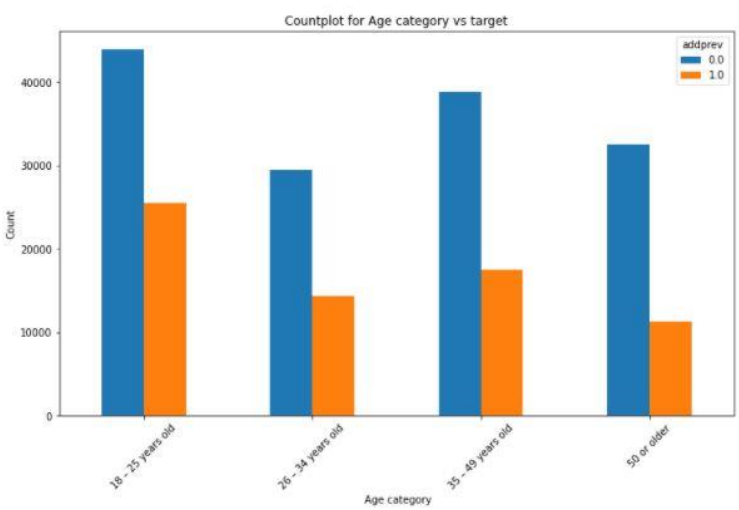
Among depressed people, 8.35% people consume marijuana more than 6 months in a year, 85.84% people consume alcohol less than 2 months in a year



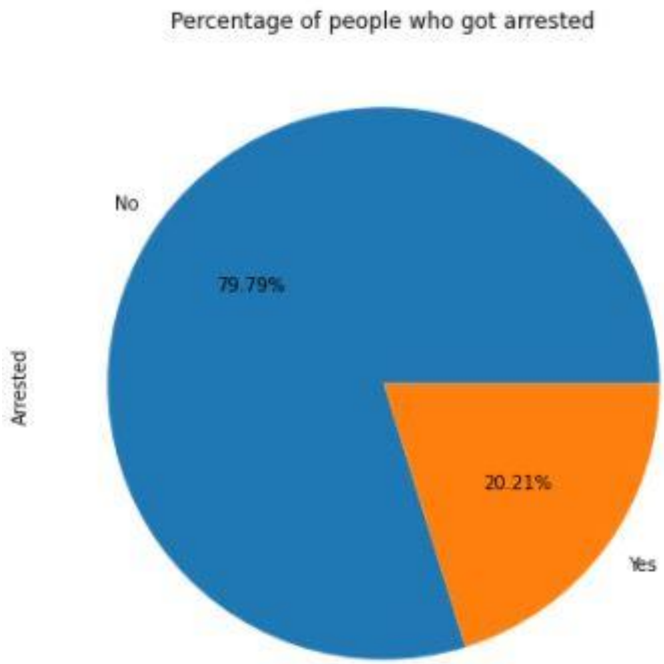


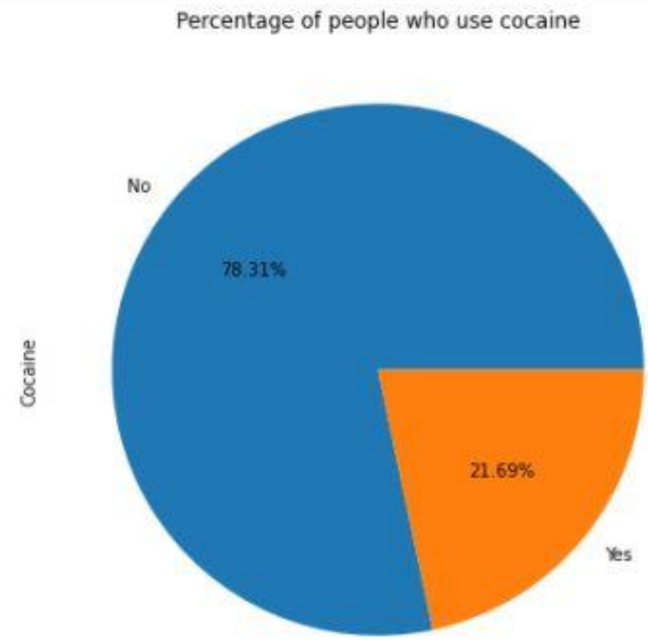
People with income less than $10000 are more prone to depression than others with high income.

People with fulltime employment are more prone to depression, followed by others (odd job workers), part-timers and unemployed



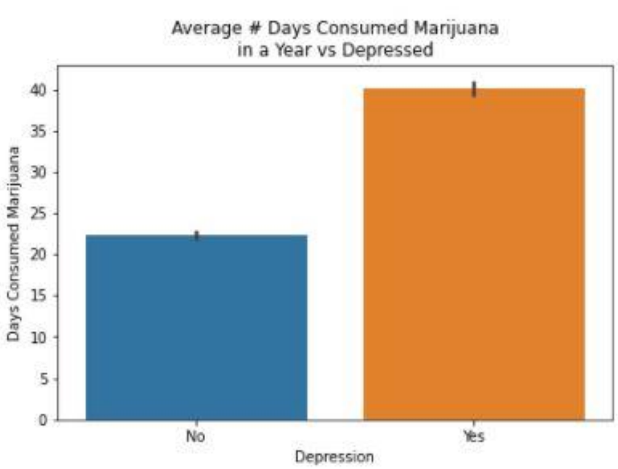
People with age 18-25 years are more prone to depression when compared to other age groups.

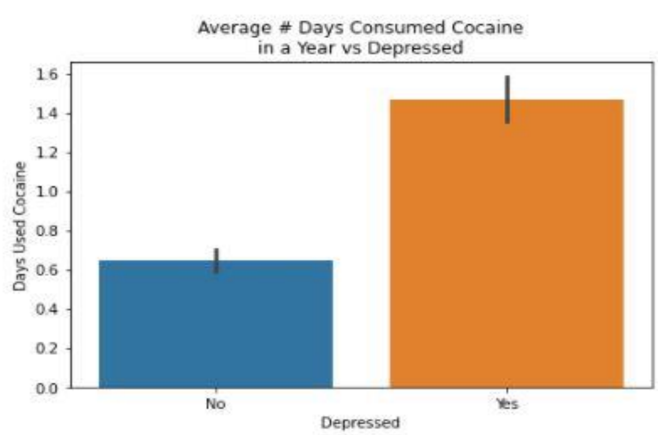




From the pie chart, we can observe that among depressed people, 79.79% are never got arrested.

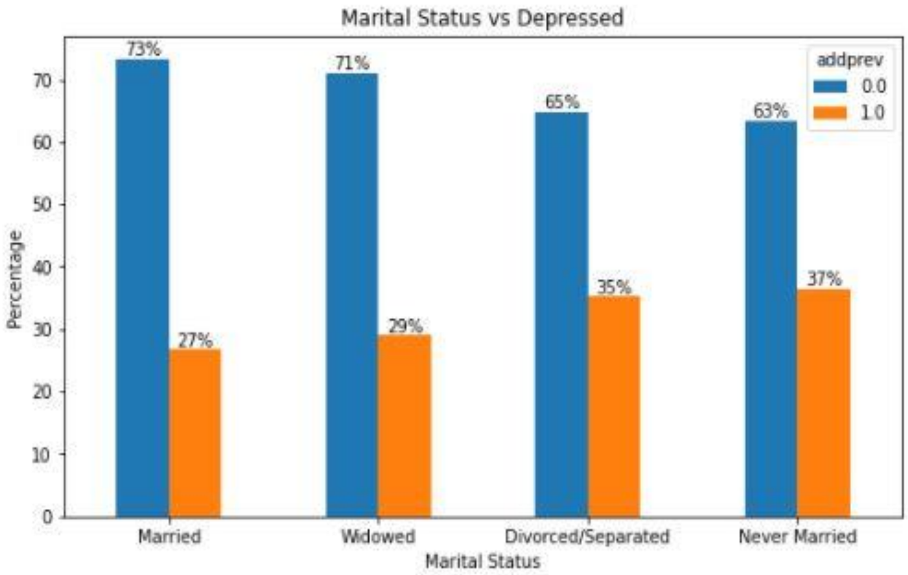
From the pie chart, we can observe that among depressed people, 78.31% never used cocaine.



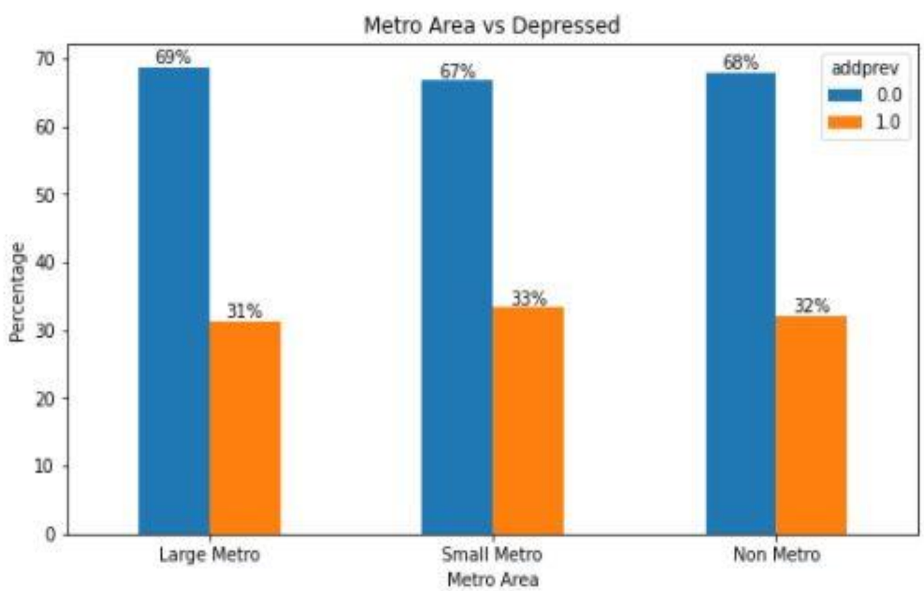


The average consumption of marijuana in a year is high for people who are depressed and not depressed.

The average consumption of cocaine in a year is high for people who are depressed and not depressed.



From the graph we can observe that, In the categories Married and widowed the ratio of not depressed to depressed is around 70:30 and in the categories Never married and divorced widowed the ratio of not depressed to depressed is around 65:35



The percentage of people among the different metro regions ratio of not depressed to depressed is around 70:30

Base Model

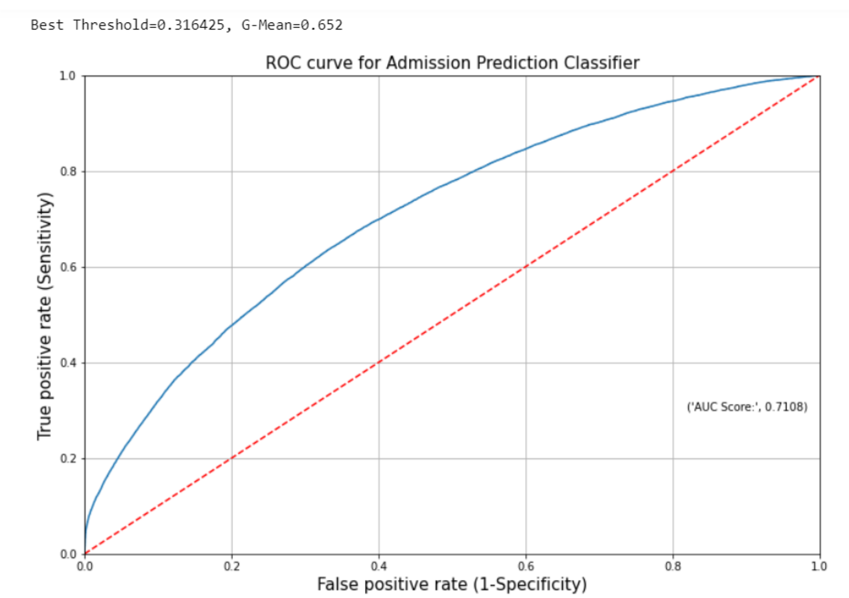
- Logistic Regression was selected for the base model because it is easier to implement, interpret, and very efficient to train.
- In the model building using interaction effect for crack and cocaine usage and for different hallucinogens
- Transforming the independent variables using standard scaler and applying the Logistic Regression.
- Base model has an accuracy of 71.38% for the train split while the test split has anaccuracy of 71.23%.

```
Results on Train data:

Accuracy on train data : 0.7138364358135293

 Confusion matrix :
 [[94543  6612]
 [36063 11910]]

 Classification report :
               precision    recall  f1-score   support

         0.0       0.72      0.93      0.82    101155
         1.0       0.64      0.25      0.36     47973

    accuracy                           0.71    149128
   macro avg       0.68      0.59      0.59    149128
weighted avg       0.70      0.71      0.67    149128
```

```
Results on Test data:

Accuracy on test data : 0.7123544874202028

 Confusion matrix :
 [[40567  2871]
 [15513  4961]]

 Classification report :
               precision    recall  f1-score   support

         0.0       0.72      0.93      0.82     43438
         1.0       0.63      0.24      0.35     20474

    accuracy                           0.71     63912
   macro avg       0.68      0.59      0.58     63912
weighted avg       0.69      0.71      0.67     63912
```

Hyper Parameters Tuning

- On top of the base model, we have developed models on Decision Tree, Random Forest, XGBoost, KNN Classifier, GaussianNB, and stacking for comparison among them.
- In order to find the best parameters for each of these models we used Grid Search CV.
- Hyper Parameter tuning for random forest: Max Depth, n Estimators, Min Samples Leaf, Min Samples Split.
- GridsearchCV best parameters obtained for XGBoost:
    - max_depth=9
    - gamma=0
    - n_estimators=100
    - learning_rate=0.1
- Also applying the threshold for the prediction as to increase the True Positive Rate (Sensitivity), taking the threshold as 0.31



Logistic Regression Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7011046438853423

 Confusion matrix :
 [[40104  3334]
 [15769  4705]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.72      0.92      0.81     43438
         1.0       0.59      0.23      0.33     20474

    accuracy                           0.70     63912
   macro avg       0.65      0.58      0.57     63912
weighted avg       0.68      0.70      0.65     63912

ROC_AUC score : 0.6841563765139417
```

Decision Tree Optimized Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7086775566403806

 Confusion matrix :
 [[40777  2661]
 [15958  4516]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.72      0.94      0.81     43438
         1.0       0.63      0.22      0.33     20474

    accuracy                           0.71     63912
   macro avg       0.67      0.58      0.57     63912
weighted avg       0.69      0.71      0.66     63912

ROC_AUC score : 0.6913411842810812
```

Random Forest Optimized Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7103517336337464

 Confusion matrix :
 [[41917  1521]
 [16991  3483]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.71      0.96      0.82     43438
         1.0       0.70      0.17      0.27     20474

    accuracy                           0.71     63912
   macro avg       0.70      0.57      0.55     63912
weighted avg       0.71      0.71      0.64     63912

ROC_AUC score : 0.6999187519744485
```

KNN Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.6843002878958568

 Confusion matrix :
 [[40021  3417]
 [16760  3714]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.70      0.92      0.80     43438
         1.0       0.52      0.18      0.27     20474

    accuracy                           0.68     63912
   macro avg       0.61      0.55      0.53     63912
weighted avg       0.65      0.68      0.63     63912

ROC_AUC score : 0.6325603844756611
```

Gaussian NB Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.6907153586180999

 Confusion matrix :
 [[38946  4492]
 [15275  5199]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.72      0.90      0.80     43438
         1.0       0.54      0.25      0.34     20474

    accuracy                           0.69     63912
   macro avg       0.63      0.58      0.57     63912
weighted avg       0.66      0.69      0.65     63912

ROC_AUC score : 0.6608048404928073
```

Adaptive Boost Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7131211666040806

 Confusion matrix :
 [[40316  3122]
 [15213  5261]]

 Classification report :
              precision    recall  f1-score   support

         0.0       0.73      0.93      0.81     43438
         1.0       0.63      0.26      0.36     20474

    accuracy                           0.71     63912
   macro avg       0.68      0.59      0.59     63912
weighted avg       0.69      0.71      0.67     63912

ROC_AUC score : 0.703509807681796
```

Gradient Boost Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7149205157091

 Confusion matrix :
 [[40848  2590]
 [15630  4844]]

 Classification report :
               precision    recall  f1-score   support

         0.0       0.72      0.94      0.82     43438
         1.0       0.65      0.24      0.35     20474

    accuracy                           0.71     63912
   macro avg       0.69      0.59      0.58     63912
weighted avg       0.70      0.71      0.67     63912

ROC_AUC score : 0.7087905346722071
```

Stacked Accuracy Scores

```
Results on Test data:

Accuracy on test data : 0.7154212041557141

 Confusion matrix :
 [[39767  3671]
 [14517  5957]]

 Classification report :
               precision    recall  f1-score   support

         0.0       0.73      0.92      0.81     43438
         1.0       0.62      0.29      0.40     20474

    accuracy                           0.72     63912
   macro avg       0.68      0.60      0.60     63912
weighted avg       0.70      0.72      0.68     63912

ROC_AUC score : 0.7129774184913008
```

XG Boost Optimized Accuracy Scores

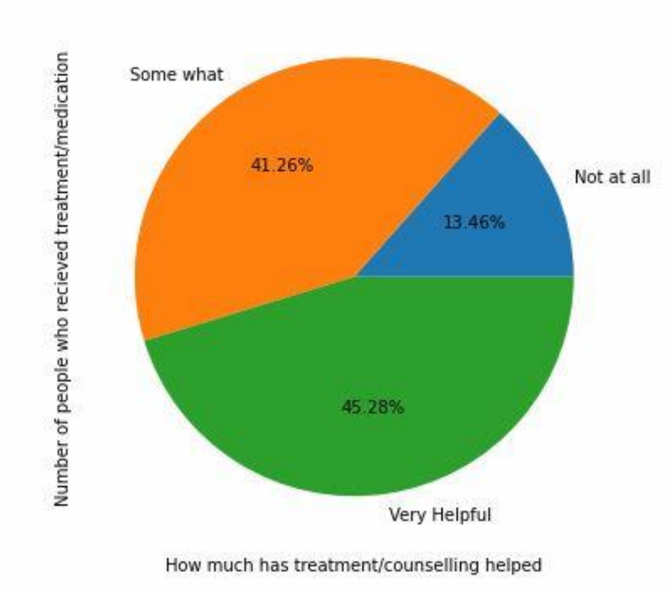
```
Results on Test data:

Accuracy on test data : 0.6508636875704094

 Confusion matrix :
 [[28219 15219]
 [ 7095 13379]]

 Classification report :
               precision    recall  f1-score   support

         0.0       0.80      0.65      0.72     43438
         1.0       0.47      0.65      0.55     20474

    accuracy                           0.65     63912
   macro avg       0.63      0.65      0.63     63912
weighted avg       0.69      0.65      0.66     63912

ROC_AUC score : 0.7107632290730679
```

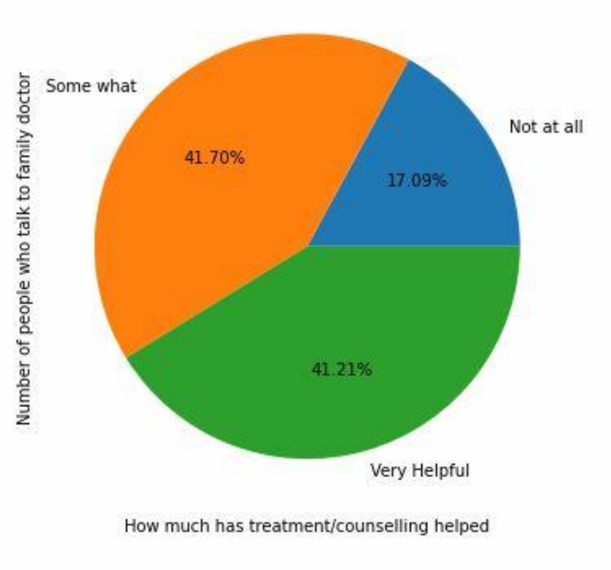
Comparison and selection of model:

	Model_Name	Train_Accuracy	Test_Accuracy	Sensitivity	Specificity
0	Logistic Regression	0.713	0.713	0.252	0.931
1	Decision Tree	0.714	0.709	0.221	0.939
2	Random Forest	0.713	0.710	0.170	0.965
3	XGBoost	0.698	0.651	0.653	0.650
4	Gaussian NB	0.690	0.691	0.254	0.897
5	KNN	0.726	0.684	0.181	0.921
6	Stack	0.770	0.715	0.291	0.915
7	Gradient Boosting	0.717	0.715	0.237	0.940
8	Adaptive Boosting	0.714	0.713	0.257	0.928

- By comparing the Train_Accuracy, Test_Accuracy, Sensitivity and Specificity we observe XG-Boost perform marginally better than other models.
- Logistic regression has low sensitivity.
- The Decision tree, Random Forest, K Nearest Neighbors, Stacking, Gradient Boosting and Adaptive Boosting has a high rate of False Negatives.
- Random Forest and K Nearest Neighbors have the least sensitivity.
- Random Forest and Gradient Boosting provides a very high True Negative Rate when compared to the other models.

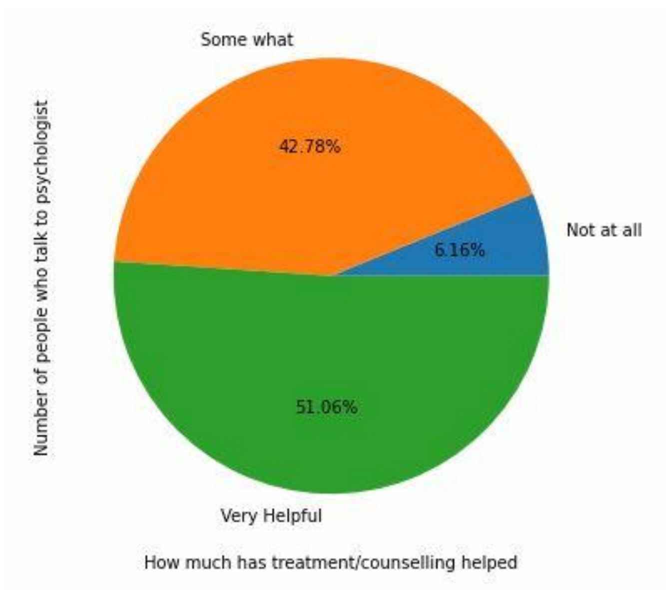
Analysis of different treatments for depression:

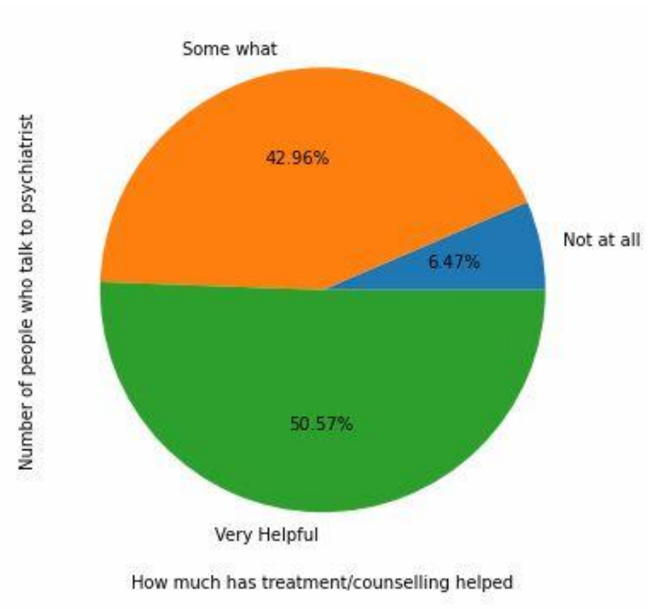




45.28% of the people who received treatment/medication for depressed feelings found to be very helpful in recovery.

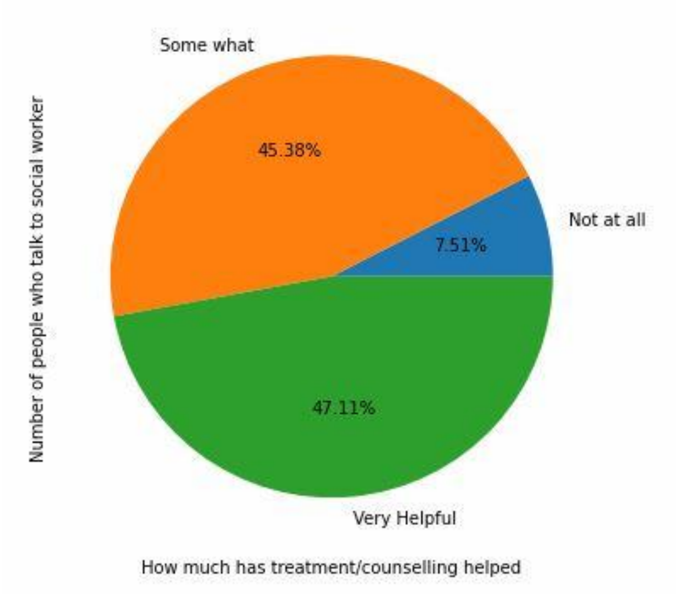
41.21% of the people who talked to general doctors/family doctors depressed feelings found to be very helpful in recovery.



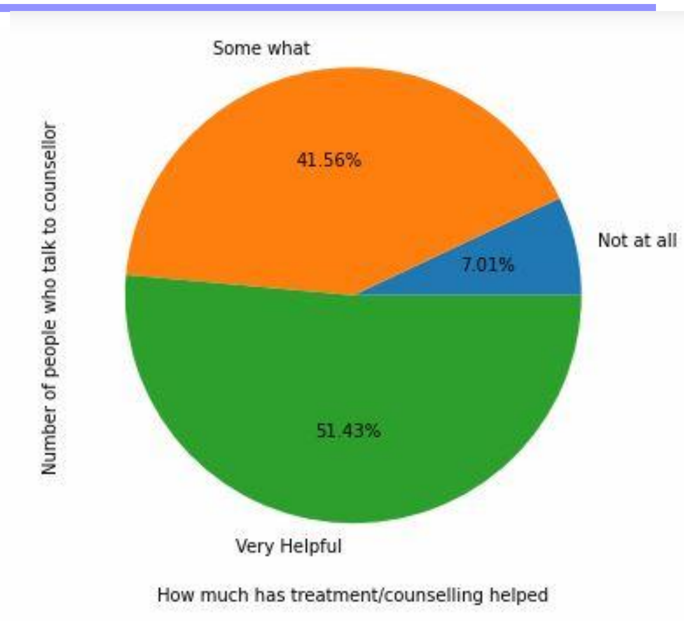


51.06% of the people who talked to psychologist about depressed feelings found to be very helpful in recovery.

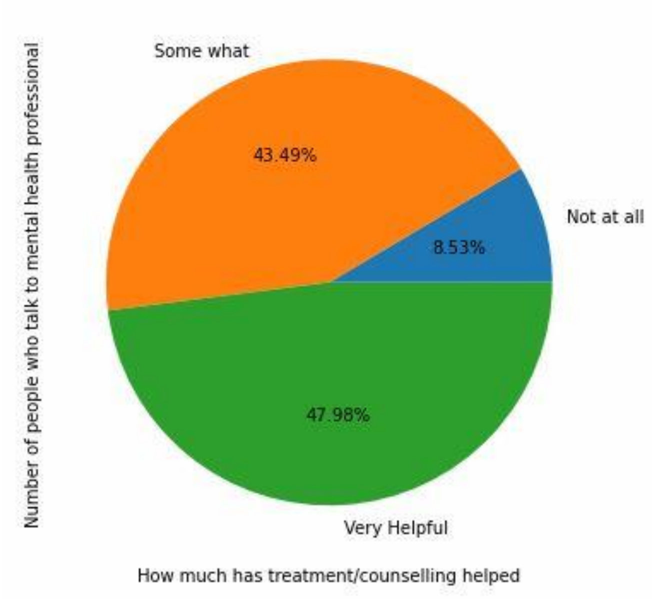
50.57% of the people who talked to psychologist about depressed feelings found to be very helpful in recovery.



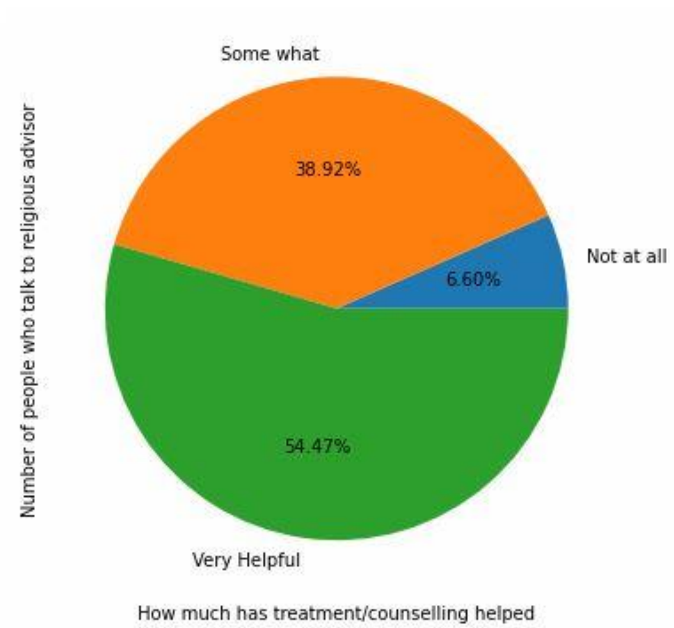
47.11% of the people who talked to social worker about depressed feelings found to be very helpful in recovery.



51.43% of the people who talked to counsellor about depressed feelings found to be very helpful in recovery.



47.98% of the people who talked to mental health professional about depressed feelings found to be very helpful in recovery.



47.98% of the people who talked to religious advisor about depressed feelings found to be very helpful in recovery.

Results and Discussion:

Stats Tests:

Hypothesis testing for categorical features:

Null hypothesis H_0 : This feature is not significant in predicting churn
Alternate Hypothesis H_a : The feature is significant in predicting churn

Stats test for categorical data

	Features	p-value
0	cigever	3.534501e-01
1	alcever	0.000000e+00
2	crkever	4.573081e-239
3	cocever	0.000000e+00
4	herever	7.720318e-229
5	methamevr	0.000000e+00
6	mjever	0.000000e+00
7	lsd	0.000000e+00

Hypothesis testing for continuous features:

Null hypothesis H_0 : This feature is not significant in predicting churn
Alternate Hypothesis H_a : The feature is significant in predicting churn

Stats test for continuous data

	Features	p-value
0	iralcfy	7.512557e-51
1	irmjfy	0.000000e+00
2	ircocfy	6.791428e-51
3	irhallucyfq	5.741773e-31
4	irinhalyfq	2.179100e-17
5	wrkdhrswk2	8.976155e-18
6	irherfy	1.848698e-23
7	irmethamyfq	2.960568e-56
8	cig30use	1.767111e-157

- As p values are less than 0.05 for all the categorical columns on performing chi-square test, therefore rejecting null hypothesis

- As p values are less than 0.05 for all the continuous columns on performing two sample t-test of independence therefore rejecting null hypothesis i.e., all the features are significantin predicting target.
- The models built ranging from the base model to the optimized models have performed fairly with models having varying accuracy as some models fit well overall while some have a low True Positive Rate while some did not perform well on test data.
- XG-Boost model is able to engage with around 65% of the depressed people as the True Positive Rate (Sensitivity) is 0.653
- Our goal is to either help people suffering from depression or prevent people who are prone to depression, it would be very bad not to approach someone who are actually depressed and cannot offer help. **To prevent False negatives, we will choose our XG-Boost Classifier because it has the highest sensitivity.**
- It is found that people with suicidal thoughts, who smoke marijuana and inhalants are more prone to depression.

Conclusion:

This study aimed at applying the machine learning techniques to predict the people who are prone to depression. Based on the analysis of the results, XG-Boost has the higher sensitivity with accuracy of 65.1%. Health institutions and hospitals can use machine learning to assess the people/patients who are prone to depression by helping them offering counselling and medication.

It is observed that, most of the people who had opted for the treatment or medication, or counselling has found to be very helpful in reducing the depressed feelings. The machine learning model helps to identify the people with drug habits and based on demographic features. It is found that people with suicidal thoughts, who smoke marijuana and inhalants are more prone to depression.

Depression is treatable with great success. The most basic kind of treatment is psychotherapy, to talk about their feelings openly and make them confident, promoting personality growth and overcome their problems.

Treatment for depression may also include antidepressants such as selective serotonin reuptake inhibitors (SSRIs), serotonin and norepinephrine reuptake inhibitors (SNRIs), tricyclic antidepressants (TCAs), monoamine oxidase inhibitors (MAOIs). Responses to antidepressants vary, and most antidepressants take 4 to 6 weeks to be fully effective. About 50% of patients respond to the first treatment, while others may have to try several different types of antidepressants before they find the best one for them.

References:

- https://www.kaggle.com/bgallamoza/national-survey-of-drug-use-and-health-20152019?select=NSDUH_2015-2019.csv
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html#sklearn.ensemble.StackingClassifier
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearch CV.html
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/