Final Project

# UNVEILING THE SHADES OF MIND: A STUDY ON DEPRESSION CLASSIFICATION

Group 08

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

## Abstract

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.
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.
Multiple models are fitted to the data to address our objective of classifying depression and the effectiveness of these models are explored by comparing different evaluation metrics. The results show that the model built with the Random Forest algorithm is the best model for prediction of presence of depression.


## Table of Contents

## List of Figures/Tables



## 1. Introduction

Depression is now identified as a serious global health issue. With newer generations battling social stigma for the betterment of their mental health, cases of depression are nowadays heard of more often than they were a few decades ago. 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. This project engages in a descriptive analysis and advanced analysis of a diverse dataset of individuals with and without depression. We utilized prior literature on the research topic to gain a good understanding of the domain and data and conducted analysis to fit multiple models and identify the best model to predict the presence of depression. This is achieved using the Python programming language and the Jupyter Notebook platform. A data product capable of efficiently classifying depression is the expected output of this project. 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.

## 2. Description of the question

This report aims to explore the capability of various machine learning algorithms to classify the presence of depression in a diverse set of patients. It targets identifying the best model for the same purpose. Medical professionals can use the identified best model to navigate treatment plans, and researchers and pharmaceutical companies with a focus on preventive medicine will also benefit from the findings of this research. However, our target audience for this project are medical professionals as we aim to enable accurate classification of depression ensuring timely intervention. This project does not address "how" to treat depression but rather "when" to treat it. The main questions that this project aims to answer are:

- Which variables have a higher impact on accurately classifying depression and to what extent?
- What machine learning models can predict depression in people, and which is the best model for the same?

## 3. 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) where 0 denotes not depressed while 1 denotes depressed. A brief description of the variables are as follows:

| Variable Name | Description | Variable Type |
|---|---|---|
| Survey_id | Survey Identification number | Nominal-Categorical |
| Ville_id | Village Identification Number | Nominal-Categorical |
| sex | Sex of the respondent; 1 = 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_investment | Indicates the absence of lasting investments; 1: No lasting investment, 0 = Presence of lasting investment. | Quantitative |
| depressed | 0 = Not depressed or 1 = depressed | Binary-Categorical |

*Table 3.1: Description of variables*

Prior to analyzing the data, it needed to be preprocessed to derive meaningful and clear interpretations. This was achieved by dropping records with missing values as the amount of missing values was not substantial.

# 4. Important results of descriptive analysis

## 4.1. Statistical Graphical Analysis of Variables Related to Depression Status

Several graphical techniques were employed under univariate, multivariate, and bivariate analyses to explore associations with respect to the depression status. Graphical plots were recognized as playing a significantly important role in identifying these associations.

Univariate analyses involved the use of histograms and box plots to visualize the distribution of numerical variables and their relationship with depression status. For bivariate analyses, scatter plots and bar charts were utilized to investigate relationships between pairs of variables and their impact on depression status.

In the multivariate context, heatmaps and pair plots were employed to examine interactions among multiple variables simultaneously. These visual representations provided valuable insights into potential correlations and associations, aiding in the understanding of how various factors relate to depression status.

### 4.1.1 Univariate Analysis

Descriptive analysis of the response variable, “depression” helped understand it better. Figure 4.1 reveals its unbalanced nature, which indicates that a higher number of individuals in the dataset do not have depression. A few different techniques, namely SMOTE, upsampling and downsampling were used to balance the data.

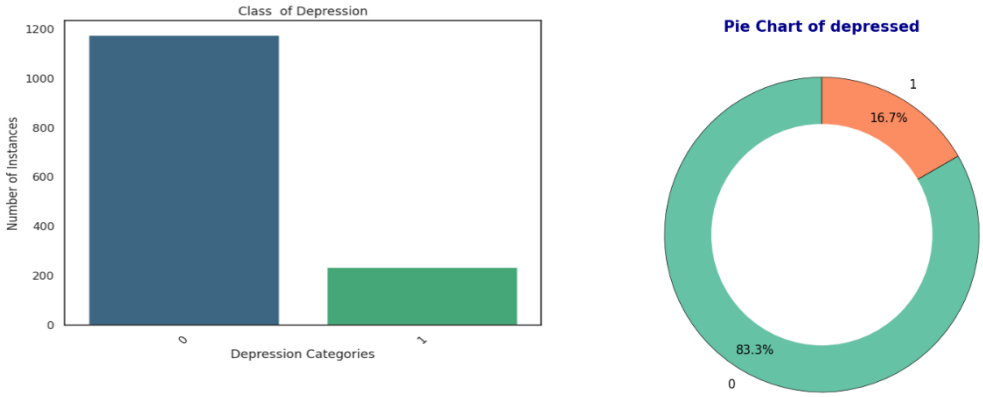


*Figure 4.1.1.1. Distribution of response variable*

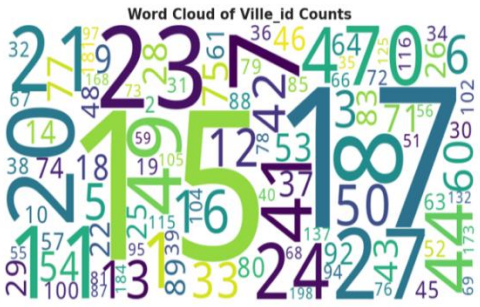


*Figure 4.1.1.2. Distribution of village ID*

A word cloud plot for the village index was generated, revealing that most individuals originate from villages 15, 17, and 27, as indicated by the prominence of these values in the visualization.

### 4.1.2. Bivariate Analysis

The distribution across different levels appears approximately same, with the shapes following a similar pattern. Peaks are observed around roughly the same values, though a higher number of observations fall under the "Not Depressed" category.

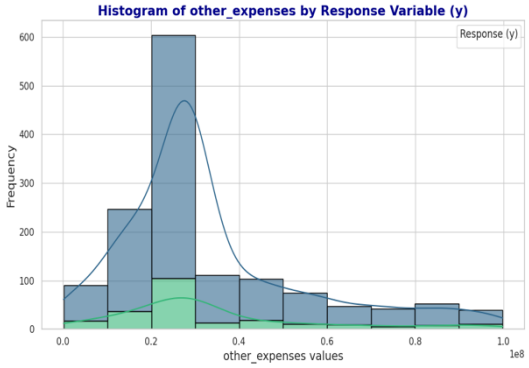


*Figure 4.1.2.1. Histogram of other expenses with response variable*

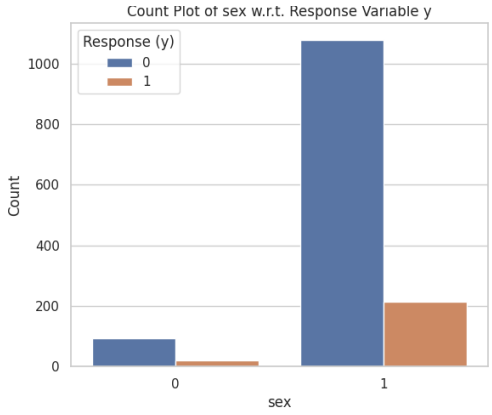


*Figure 4.1.2.2. Multiple Bar chart of other Gender with response variable*

It was observed that the majority of females fall under the "Not Depressed" category. Males also follow a similar pattern, indicating a lower likelihood of depression among them.

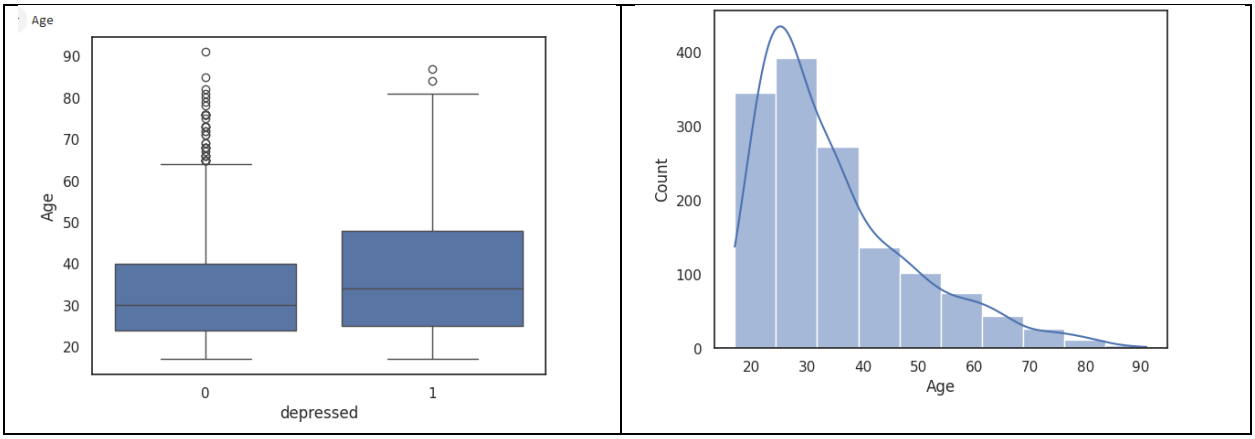


*Figure 4.1.2.3. Box Plot and Histogram of Age with response variable*

The age distribution appears skewed, with a peak concentrated around the lower age values. It was observed that most individuals in the "Depressed" category tend to be older, whereas many outliers in the "Not Depressed" category are among older individuals. However, the majority of "Not Depressed" individuals are around 30 years of age or younger

### 4.1.3. Multivariate Analysis

#### 4.1.3.1 Association Among the Categorical Variables

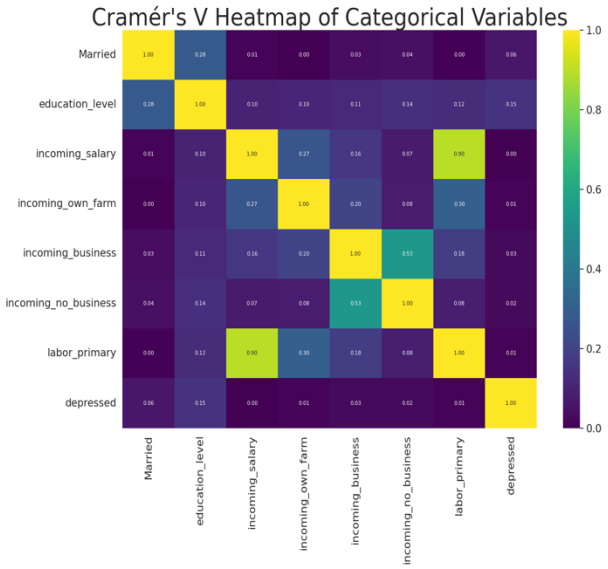


*Figure 4.1.3.1. Cramer's V Heatmap of Categorical Variables*

Only the categorical variables were considered for this analysis, where Cramér's V was employed to assess the strength of association among them. A heatmap was plotted to visually represent these associations. It was observed that none of the categorical variables were significantly associated with the response variable, depression status.

However, a significantly higher statistic value was noted between two specific categorical variables labor_primary.and incoming_own. Given that labor_primary pertains to labor status within itself, there was a consideration of removing this variable from further analyses to ensure the robustness of the findings.

#### 4.1.3.2 Association Among the Numerical Variables

The Pearson correlation coefficient was utilized to assess the associations among the numerical variables, and heatmaps were employed for visual representation of these correlations. It was observed that there was a high correlation (0.78) between the number of family members and the number of children. Since the number of family members inherently includes the number of children, this redundancy led to the decision to remove the variable number_children from the analysis, allowing for a more streamlined assessment of the data.

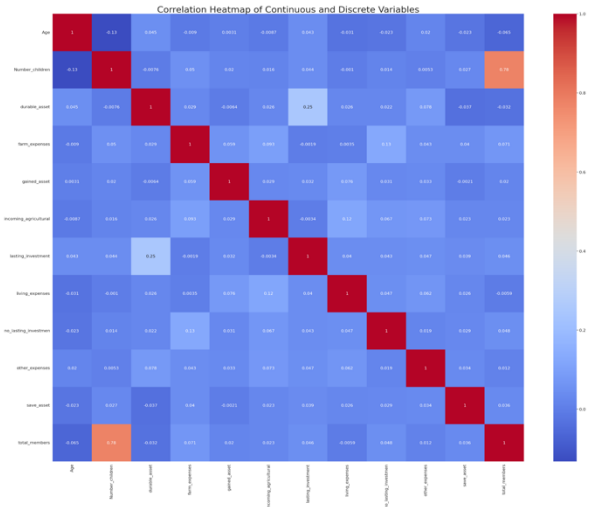


*Figure 4.1.3.2. Pearson Correlation Heatmap of Numerical Variables*

### 4.2 Statistical Numerical Analysis of Variables Related to Depression Status

Several numerical tests were conducted on the dataset to evaluate the significance of the numerical variables in relation to the categorical response variable, depression status. The Kruskal-Wallis test was performed at a 5% significance level, revealing that 'Age' and 'total_members' were significantly associated with the categorical response variable of depression status.

To assess the significance of categorical variables with respect to depression status, the Chi-Square test was utilized. At the same 5% significance level, it was found that none of the categorical variables exhibited a significant association with depression status.

Furthermore, Cramér's V statistics were calculated, reinforcing the findings from the Chi-Square test, as none of the variables demonstrated a significant association with depression status. This analysis indicates that while certain numerical variables may have relevance, categorical variables did not show significant relationships in this context.

### 4.3 Clustering

Since the dataset consists of both categorical and quantitative variables, k-prototypes clustering was applied. However, the silhouette score was found to be very low, with a value close to zero (0.27). Upon examining the cluster distributions, significant imbalance was observed: Cluster 0 contained 163 observations, Cluster 1 had 199 observations, Cluster 2 had 612 observations, and Cluster 3 had 153 observations.

Additionally, KMeans and k-prototypes clustering algorithms were modified using custom distance measures to accommodate the mixed data types. However, both approaches also produced low silhouette scores, with values of 0.28 and 0.05, respectively. Consequently, further analysis with clustering methods was not pursued due to the low silhouette scores and the presence of imbalanced cluster sizes.

## 5. Important results of advanced analysis

Upon completion of preprocessing, we fitted the above models to the data. The data set was splitted into training (80% of original data) and test (20%) tests in the pre-processing stage and the above-mentioned models were fitted on the training data and tested on the test data.

### 5.1 Selected models and justification of choice

**1. Logistic Regression:** A simple, interpretable technique that provides probabilities of class membership makes it useful for understanding the likelihood of depression incidence based on various features.

**2. Support Vector Machine (SVM):** SVMs can find optimal hyperplanes to separate classes, potentially capturing intricate decision boundaries between people that are depressed and not depressed.

**3. Decision Tree:** Decision trees are intuitive and capture non-linear relationships between features and the target variable, making them suitable for prediction tasks like classifying depression occurrence. They can handle both numerical and categorical data.

**4. Random Forest:** Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and generalization and is robust to overfitting.

**5. XGBoost:** A powerful gradient boosting algorithm, it is known for its performance and scalability. Effectively handling imbalanced datasets and it captures complex interactions

between features, making it well-suited for depression classification where class distribution may be highly skewed.

**6. k-Nearest Neighbours (kNN):** kNN is simple and easy to understand, thereby well suited for initial exploration of the
data and potentially capturing localized patterns relevant to depression classification.

**7. AdaBoost:** Adaboost is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. Adapting to the complexity of the dataset, it potentially improves predictive performance when data is highly imbalanced.

**8. Gradient Boosting:** This method builds a series of decision trees sequentially, with each tree correcting errors made by previous ones.

**9. MLP (Multi-Layer Perceptron):** MLP is an artificial neural network capable of learning complex relationships between input features and the target variable. It's helpful when there are non-linear relationships and interactions between multiple variables.

**10. Naive Bayes Classifier:** Naive Bayes is a probabilistic classifier which applies Bayes' theorem with a "naive" assumption of independence between features. It is computationally efficient and often effective for many classification tasks.

### 5.2 Handling Class Imbalance and Evaluation of Models on Data Set after applying different sampling techniques

We observed a substantial class imbalance within our dataset, with approximately 83.32% of cases not having depression within the studied population. To address this imbalance, we applied several techniques which were SMOTE (Synthetic Minority Over-sampling Technique), upsampling and downsampling. The same set of algorithms were then used to fit various models, and these were evaluated using the previously mentioned metrics.

#### 5.2.1 Smote dataset

We fitted the same set of models to SMOTE data. The table below presents the performance metrics obtained from these models with hyperparameter tuning on SMOTE data:

| | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Log Reg | 0.546 | 0.501 | 0.501 | 0.457 |
| SVM | 0.805 | 0.416 | 0.481 | 0.446 |
| Decision tree | 0.624 | 0.497 | 0.495 | 0.481 |
| Random forest | 0.759 | 0.518 | 0.515 | 0.515 |
| XGBoost | 0.741 | 0.522 | 0.522 | 0.522 |
| KNN | 0.553 | 0.473 | 0.453 | 0.437 |
| AdaBoost | 0.688 | 0.492 | 0.49 | 0.489 |
| Gradient Boost | 0.752 | 0.502 | 0.502 | 0.501 |
| MLP | 0.688 | 0.506 | 0.507 | 0.503 |
| Naive Bayes | 0.415 | 0.476 | 0.458 | 0.378 |

*Table 5.2.1. Evaluation metrics of models fitted on SMOTE data*

The support vector machine and random forest models emerged as the top performers with evaluation metrics, accuracy: 80.5%, precision: 41.6%, F1-score: 44.6% and recall: 48.1% and accuracy: 75.9%, precision: 51.8%, F1-score: 51.5% and recall: 51.5% respectively.

### 5.2.2 Upsampled dataset

We fitted the same set of models to upsampled data. The table below presents the performance metrics obtained from these models with hyperparameter tuning on upsampled data:

| | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Log Reg | 0.521 | 0.483 | 0.469 | 0.434 |
| SVM | 0.837 | 0.418 | 0.5 | 0.456 |
| Decision tree | 0.727 | 0.453 | 0.461 | 0.456 |
| Random forest | 0.823 | 0.502 | 0.5 | 0.47 |
| XGBoost | 0.783 | 0.489 | 0.494 | 0.483 |
| KNN | 0.606 | 0.473 | 0.459 | 0.454 |
| AdaBoost | 0.642 | 0.492 | 0.488 | 0.481 |
| Gradient Boost | 0.791 | 0.497 | 0.499 | 0.487 |
| MLP | 0.745 | 0.553 | 0.559 | 0.555 |
| Naive Bayes | 0.638 | 0.485 | 0.478 | 0.473 |

*Table 5.2.2. Evaluation metrics of models fitted on upsampled data*

The support vector machine and random forest models emerged as the top performers with evaluation metrics, accuracy: 83.7%, precision: 41.8%, F1-score: 45.6% and recall: 50% and accuracy: 82.3%, precision: 50.2%, F1-score: 47% and recall: 50% respectively.

### 5.2.3 Downsampled dataset

We fitted the same set of models to downsampled data. The table below presents the performance metrics obtained from these models with hyperparameter tuning on downsampled data:

| | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Log Reg | 0.504 | 0.501 | 0.502 | 0.438 |
| SVM | 0.805 | 0.416 | 0.481 | 0.446 |
| Decision tree | 0.337 | 0.505 | 0.508 | 0.332 |
| Random forest | 0.454 | 0.514 | 0.525 | 0.418 |
| XGBoost | 0.468 | 0.504 | 0.507 | 0.422 |
| KNN | 0.536 | 0.55 | 0.591 | 0.484 |
| AdaBoost | 0.422 | 0.503 | 0.506 | 0.394 |
| Gradient Boost | 0.514 | 0.538 | 0.57 | 0.466 |
| MLP | 0.532 | 0.515 | 0.528 | 0.461 |
| Naive Bayes | 0.415 | 0.476 | 0.458 | 0.378 |

*Table 5.2.3. Evaluation metrics of models fitted on downsampled data*

The support vector machine model emerged as the top performer with evaluation metrics, accuracy: 80.5%, precision: 41.6%, F1-score: 44.6% and recall: 48.1%.

### 5.3 Evaluation of Classification Models on Original Training Set considering outlier and regular observations clusters

During the exploratory data analysis (EDA), 30.1% of the data points were identified as multivariate outliers using an Isolation Forest with the contamination parameter set to "auto." Given the substantial proportion of outliers, outright removal was deemed impractical due to potential information loss. Instead, the dataset was divided into two subsets: inliers and outliers. Ten different models were then trained and evaluated separately on the training and test sets within each cluster to assess the overall model performance. Among the models, the Random Forest classifier achieved the best performance, yielding an accuracy of 85.5%, precision of 80.7%, recall of 85.5%, and F1-score of 80%, and was thus selected as the optimal model for this dataset.

The table below presents the performance metrics obtained from the models fitted on Original Training Set considering outlier and regular observations clusters:

| | Outlier cluster - Test Accuracy | Regular obvs - Test Accuracy | Overall | | | |
|---|---|---|---|---|---|---|
| | | | Overall Accuracy | Overall Precision | Overall Recall | Overall F1 |
| Log Reg | 0.859 | 0.853 | 0.855 | 0.730 | 0.855 | 0.788 |
| SVM | 0.859 | 0.853 | 0.855 | 0.730 | 0.855 | 0.876 |
| Decision tree | 0.694 | 0.726 | 0.716 | 0.752 | 0.716 | 0.733 |
| Random forest | 0.859 | 0.853 | 0.855 | 0.807 | 0.855 | 0.800 |
| XGBoost | 0.812 | 0.812 | 0.812 | 0.747 | 0.812 | 0.775 |
| KNN | 0.8 | 0.838 | 0.826 | 0.743 | 0.826 | 0.779 |
| AdaBoost | 0.776 | 0.843 | 0.823 | 0.762 | 0.823 | 0.786 |
| Gradient Boost | 0.788 | 0.848 | 0.83 | 0.727 | 0.83 | 0.775 |
| MLP | 0.671 | 0.492 | 0.546 | 0.757 | 0.546 | 0.613 |
| Naive Bayes | 0.824 | 0.848 | 0.84 | 0.729 | 0.84 | 0.78 |

*Table 5.3.1. Evaluation metrics of models fitted on original data*

### 5.4 Selection of best model

We consider the model with highest accuracy, F1 Score, precision and recall value as our best model. The Random Forest classifier with an accuracy of 85.5%, precision of 80.7%, recall of 85.5%, and F1-score of 80%, was thus selected as the optimal model.

### 5.5 Feature importance of best model

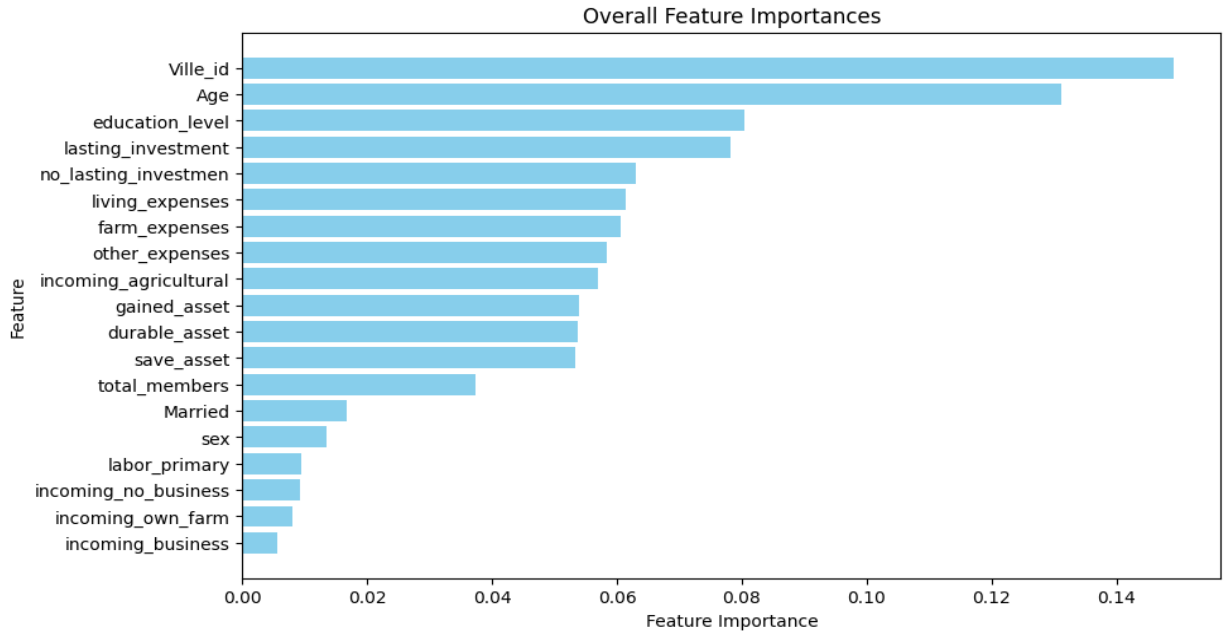


*Figure 5.5.1. Feature importance of best model*

Feature importance in descending order of the variables included in the optimal model can be seen from figure 5.5.1 above. There are 19 features in this model.

When fitting the model to only the top 10 features there is a slight drop in the accuracy, therefore we considered the model with all 19 variables as the optimal model.

## 6. Outputs

The main output of our study is the data product which is a website that interested parties can use for classification of depression status.

## 7. Discussion and conclusions

Initial examination of the data revealed class imbalance in the response variable “depression”. We addressed this issue by implementing SMOTE, upsampling and downsampling. The advanced analysis was conducted on the original dataset as well as the balanced datasets. Further models were also fitted to the data while considering the existence of an outlier cluster along with a regular observations cluster. Various models were constructed, with the optimal model being selected by comparing accuracy, f1-score, precision and recall among other evaluation metrics.

The random forest model fitted on the original data emerged as the top performer based on the following metrics: 85.5%, precision of 80.7%, recall of 85.5%, and F1-score of 80%.

- Accuracy ( 85.5%): This high accuracy indicates that the model correctly predicts the outcome for 85.5% of instances in the dataset.

- Precision (80.7%): This high precision indicates that when the model predicts a positive outcome, it is correct 80.7% of the time and so the model has a relatively low rate of false positives.
- Recall (85.5%): This high recall indicates that the model correctly identifies 85.5% of all actual positive cases in the dataset. That is the model is effective at capturing most of the true positives.
- F1-score (80%): The F1-score combines precision and recall into a single metric, balancing the two. An F1-score of 80% reflects a solid trade-off between precision and recall, demonstrating that the model performs well overall in identifying true positives while controlling false positives.

## 8. Appendix

Link to code:
https://drive.google.com/drive/folders/1bLj8_U2ND2qYBi5cqmCpFq7SuAkzcauF?usp=sharing