Early Diagnosis of Depression

1. Project Definition

1.1 Project Overview

This project aims to predict whether an individual is suffering from depression using machine learning models. The dataset, sourced from Zindi[1], contains survey responses related to mental health in East Africa. By building a predictive model, we hope to assist in the early diagnosis of depression, which can lead to timely and effective interventions. The project includes a web interface that allows users to input relevant data and receive immediate predictions regarding their mental health status.

1.2 Problem Statement

Depression is a mental disorder that negatively affects a person's behaviors, actions, and thoughts. It leads to loss of interest, appetite, sleep problems, and fatigue. Moreover, depression can exacerbate physical symptoms and complicate chronic diseases such as diabetes, hypertension, cancer, stroke, and heart attack. According to the World Health Organization (WHO), 3.8% of the world's population is affected by depression, including 5% of adults. In Saudi Arabia, studies indicate that the prevalence of depression ranges from 17% to 46% of the population. Given the significant impact of depression on individuals and public health, early diagnosis is crucial. This project aims to implement a machine learning model to predict depression based on survey data, aiding in early detection and intervention.

1.3 Metrics

To evaluate the performance of the machine learning models used in this project, we will consider the following metrics:

- **Accuracy:** Measures the proportion of correctly predicted instances out of the total instances. It is a basic metric that provides an overall sense of how well the model is performing.
    - Formula:

      $$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- **ROC AUC Score:** Represents the area under the Receiver Operating Characteristic (ROC) curve. It measures the model's ability to distinguish between classes and is useful for evaluating models on imbalanced datasets.
    - Interpretation: A score of 0.5 indicates a model with no discriminative power, while a score of 1.0 indicates perfect discrimination.

[1] **Zindi Africa, "Busara Mental Health Prediction Challenge,"** [Online]. Available: https://zindi.africa/competitions/busara-mental-health-prediction-challenge/data. [Accessed: 08-Jun-2024].

2. Analysis

2.1 Data Exploration and Visualization

This project utilized a public online dataset available at Zindi, provided by the Busara Center for Behavioral Economics. The data is a collection of survey responses related to mental health in East Africa. It includes information about respondents' demographic characteristics, socio-economic status, and mental health status.

The dataset consists of 1143 records with 75 attributes, including the target class. The target variable is the participant's mental health status, a binary variable indicating whether the participant has a mental health condition or not. Out of 1143 participants, 193 have a mental health condition.

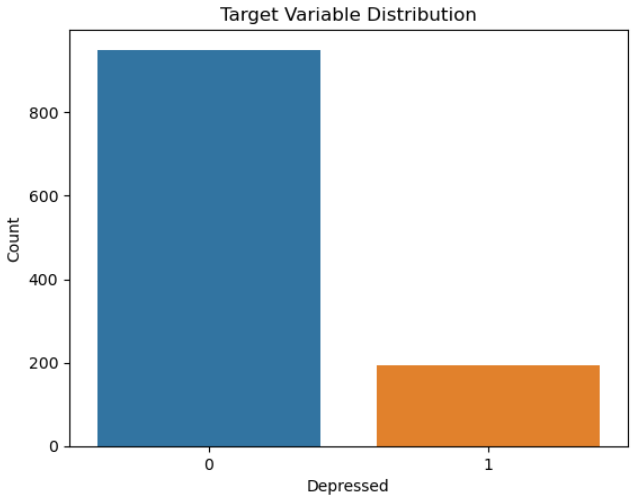

Figure 1: Target Variable Distribution

Figure 1 above illustrates a significant class imbalance in the dataset. The majority of participants are not depressed, while a smaller portion of participants are depressed. This imbalance needs to be considered when building and evaluating the machine learning model, as it may affect the model's performance and predictions.

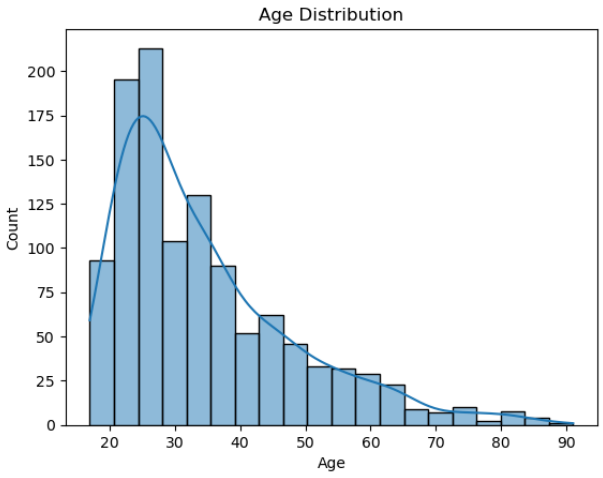

Figure 2: Age Distribution

Figure 2 depicts the age distribution of participants in the dataset. The histogram reveals that most participants are in their early 20s, with the highest concentration around age 23. While the ages of participants span from 20 to approximately 90, the majority are under 50. This skew towards younger individuals is a key observation, as it may influence the analysis and the predictions of the model related to depression.

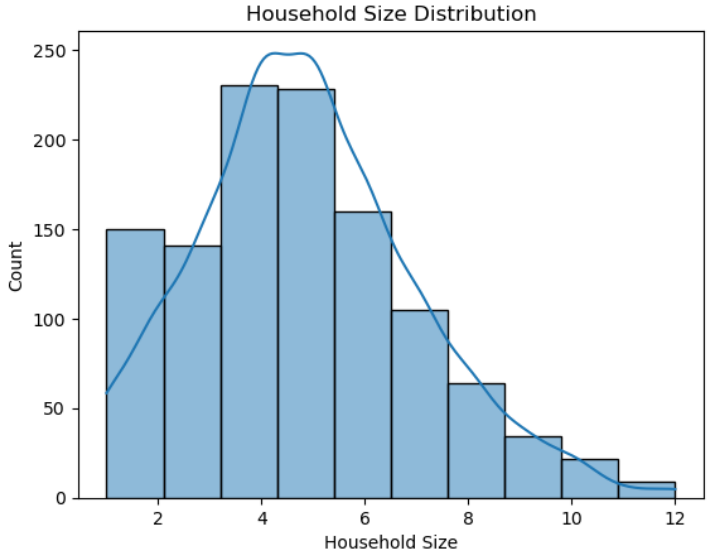

Figure 3: Household Size Distribution

Figure 3 illustrates the distribution of household sizes among participants in the dataset. The histogram shows that most households have 4 to 5 members, with the highest concentration around 4 members. Household sizes range from 1 to 12 members, but the majority of households fall between 2 and 8 members. This common household size may influence how household size relates to depression in the analysis, as the dataset is centered around households with 4 to 5 members.

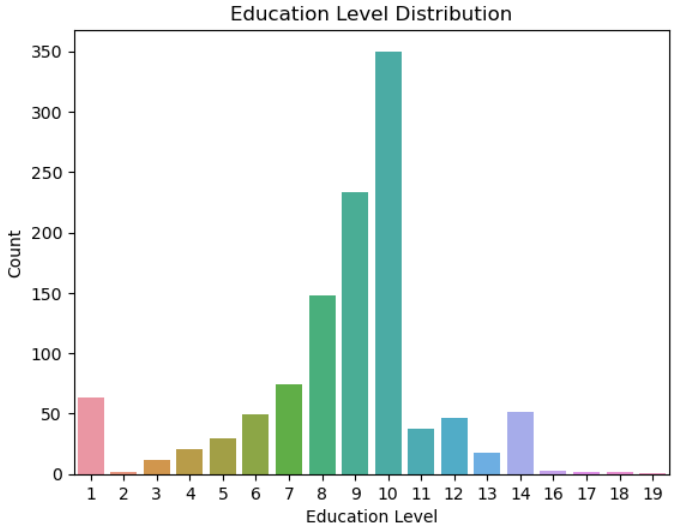

Figure 4: Education Level Distribution

Figure 4 presents the distribution of education levels among participants in the dataset. The bar chart indicates that the most common education level is 10, with more than 350 participants at this level. While education levels range from 1 to 19, the majority of participants are clustered between levels 6 and 12. This peak at education level 10 is significantly higher than other levels, which may influence how education relates to depression in the analysis.

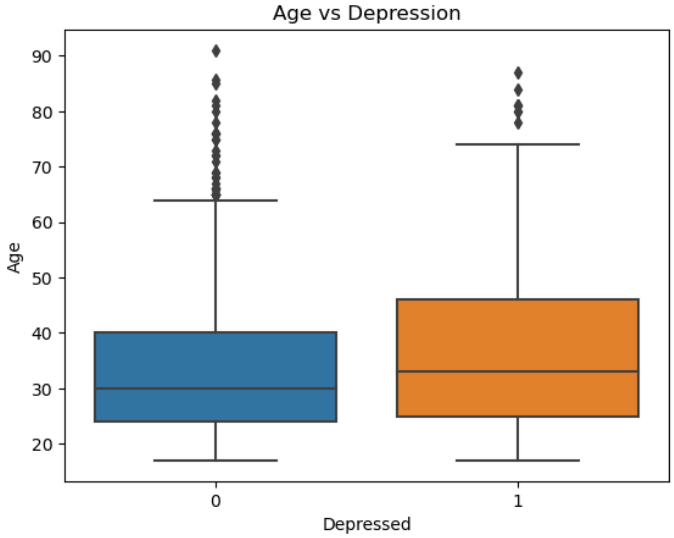

Figure 5: Age VS Depression

Figure 5 compares the age distribution between participants who are not depressed (0) and those who are depressed (1) using a boxplot. For non-depressed participants (0), the median age is around 30 years, with most participants falling between 20 and 40 years old. There are several outliers above the age of 60. For depressed participants (1), the median age is slightly higher, around 35 years, with most participants aged between 25 and 45. Similarly, there are outliers above the age of 60. Furthermore, participants with depression tend to be slightly older than those without, and the age range is broader for depressed individuals. This suggests that age might play a role in the likelihood of experiencing depression.

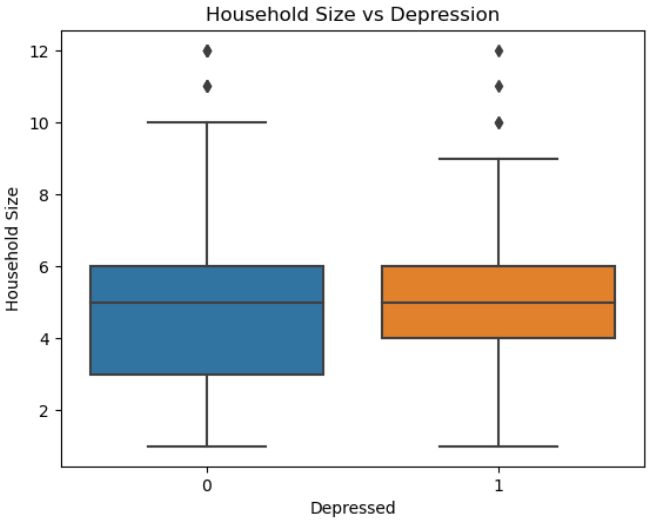

Figure 6: Household Size VS Depression

Figure 6 compares the household size distribution between participants who are not depressed (0) and those who are depressed (1) using a boxplot. For non-depressed participants (0), the median household size is 5 members, with most households having between 4 and 6 members. There are outliers with household sizes above 10. For depressed participants (1), the median household size is slightly smaller, around 4 members, with most households having between 4 and 6 members. Similarly, there are outliers with household sizes above 10. Moreover, both groups have similar household size distributions, although non-depressed participants tend to have slightly larger households on average. Household size appears to have a minor impact on depression status.

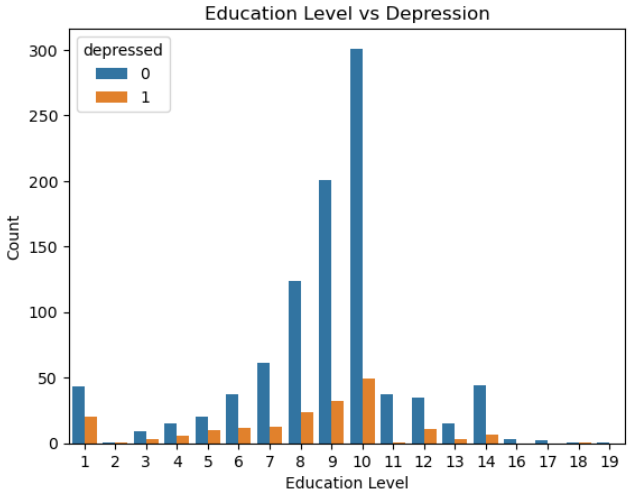

Figure 7: Education Level VS Depression

Figure 7 compares the education level distribution between participants who are not depressed (0) and those who are depressed (1) using a bar chart. Education level 10 has the highest count for both non-depressed and depressed participants, with a significantly larger number of non-depressed participants. While most education levels have a higher count of non-depressed participants, there is a noticeable presence of depressed participants across various education levels, especially around levels 8, 9, and 10. This distribution indicates that education level may influence depression, but other factors likely play a significant role as well.

3. Methodology

3.1 Data Preprocessing

3.1.1 Handling Missing Values

To prepare the dataset for prediction, several preprocessing steps were taken to handle missing values. First, unnecessary columns such as '*surveyid'*, '*village*', '*survey_date'*, and '*day_of_week*' were dropped, reducing the dataset from 1144 records with 75 attributes to 1143 records with 71 attributes. After that, columns with missing values greater than or equal to 25% were removed, further reducing the dataset to 50 attributes. Finally, rows with missing values exceeding 2% of their total columns were dropped, resulting in a final dataset of 1094 records with 50 attributes. After these steps, the dataset was free of any remaining missing values.

3.1.2 Normalization

To ensure the features were on a similar scale and to improve the performance of machine learning models, the dataset underwent normalization using the Robust Scaler. This scaling method was chosen due to its robustness to outliers. The dataset was split into two parts: one containing the binary and categorical features and the other containing the continuous features. The continuous features were normalized while the binary and categorical features were retained in their original form. After normalization, the two parts were combined back into a single dataset, maintaining a consistent format for further analysis.

3.1.3 Feature Selection

For selecting the most important features, the SelectKBest method with the f_classif score function was used. This process involved iteratively selecting the top k features and evaluating model performance. Ultimately, 24 features were identified as the most important, and these were saved for further analysis.

3.2 Implementation and Refinement

3.2.1 Model Building

The process of model building began with splitting the dataset into training and testing sets using a 20% test size with a random state value of 123 for reproducibility. The resulting split included 728 non-depressed and 147 depressed participants in the training set, and 183 non-depressed and 36 depressed participants in the testing set.

3.2.2 Undersampling using NearMiss

Due to the class imbalance in the dataset, NearMiss undersampling was applied to balance the classes in both training and testing sets. This technique brought the number of non-depressed and depressed participants to 147 each in the training set, and 36 each in the testing set.

3.2.3 Modeling

Three algorithms were utilized to build the predictive models: Support Vector Machine (SVM), Decision Tree (DT), and k-Nearest Neighbor (k-NN). To enhance the performance of these models, GridSearchCV was used for hyperparameter tuning. This technique performs an exhaustive search over specified parameter values for an estimator, enabling the identification of the best combination of hyperparameters to optimize model performance.

4. Results

4.1 Model Evaluation and Validation

This section compares the performance of three different machine learning models: K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM). The results are evaluated based on the training score, confusion matrix, accuracy, and ROC AUC.

Table 1: Optimized Models Evaluation Results

Algorithm	**Training Accuracy**	**Testing Accuracy**	**ROC-AUC**
k-NN	84.35%	81.94%	81.94%
DT	86.05%	77.78%	77.78%
SVM	88.10%	83.33%	83.33%

As show in Table 1, the KNN model shows a good performance with an accuracy of 81.94% on the test set. It correctly identified 33 non-depressed and 26 depressed cases, with 13 misclassifications (3 false positives and 10 false negatives). The train score of 84.35% indicates that the model performs well on the training data as well. The ROC AUC score of 81.94% is consistent with the accuracy, indicating a reliable performance.

The DT model has a slightly lower accuracy of 77.78% compared to the KNN model. It correctly predicted 31 non-depressed and 25 depressed cases, with 16 misclassifications (5 false positives and 11 false negatives). The train score of 86.05% suggests the model fits the training data well. The ROC AUC score of 77.78% aligns with the accuracy, indicating a somewhat lower performance compared to KNN and SVM.

The SVM model achieved the highest accuracy of 83.33% on the test set. It correctly identified 32 non-depressed and 28 depressed cases, with 12 misclassifications (4 false positives and 8 false negatives). The train score of 88.10% indicates that the model performs the best on the training data among the three models. The ROC AUC score of 83.33% is the highest, reflecting the model's strong performance in distinguishing between depressed and non-depressed cases.

Among the three models evaluated, the SVM model demonstrates the best performance with the highest accuracy (83.33%) and ROC AUC score (83.33%). The KNN model also performs well with an accuracy of 81.94% and ROC AUC of 81.94%. The DT model, while having a good train score, has the lowest test accuracy (77.78%) and ROC AUC score (77.78%) among the three models. These results suggest that the SVM model is the most effective for predicting depression in this dataset.

4.2 Justification

The final results of the three machine learning models: k-NN, DT, and SVM are compared in detail to explain why certain models, parameters, or techniques performed better than others.

SVM:

- **Hyperparameters:** The SVM model was tuned using GridSearchCV, optimizing parameters like C (regularization parameter) and gamma (kernel coefficient). The best parameters found were C=4 and gamma=0.1, which helped the model generalize better on the test set.
- **Justification:** SVM's effectiveness lies in its ability to find the optimal hyperplane that maximizes the margin between different classes. This makes it particularly suitable for high-dimensional spaces and datasets with clear margins of separation.

k-NN:

- **Hyperparameters:** GridSearchCV was used to tune the n_neighbors and metric parameters. The best results were obtained with n_neighbors=7.
- **Justification:** KNN's performance can be attributed to its simplicity and effectiveness in capturing local patterns in the data. However, its performance can degrade with imbalanced data or higher-dimensional spaces, which might explain its slightly lower performance compared to SVM.

DT:

- **Hyperparameters:** The model was tuned using GridSearchCV with parameters like max_features, criterion, splitter, and max_depth. The best model used max_features='auto' and max_depth=5.
- **Justification:** Decision Trees are prone to overfitting, especially with imbalanced datasets. The lower performance indicates that the model might have overfitted to the training data, failing to generalize well on unseen data.

5. Conclusion

5.1 Reflection

In this project, I have tackled the problem of predicting depression using a dataset from Zindi. The end-to-end solution involved data preprocessing, feature selection, model training, and evaluation. I have evaluated three machine learning models: K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM). The SVM model demonstrated the best performance with an accuracy of 83.33% and an ROC AUC score of 83.33%.
One interesting aspect of this project was handling the class imbalance in the dataset, which significantly influenced model performance. Implementing feature selection using the wrapping technique was also challenging but crucial for improving model accuracy and interpretability.

5.2 Improvement

To further improve the experiment, several suggestions can be made for future research:

- **Advanced Imbalance Handling**:
  While NearMiss undersampling was effective, exploring other techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ensemble methods like Balanced Random Forest could provide better performance by creating synthetic samples or combining multiple models to handle class imbalance more effectively.
- **Feature Engineering**:
  Additional feature engineering could be conducted to create new features that capture more relevant information from the existing data. For example, combining related features or creating interaction terms could potentially enhance the models' predictive power.
- **Model Ensemble**:
  Implementing an ensemble approach by combining the strengths of multiple models (e.g., stacking, boosting) could lead to improved overall performance. Ensembles often provide better generalization by leveraging the diversity of different models.

- **Cross-Validation**:
  Although Stratified K-Fold cross-validation was used during hyperparameter tuning, employing nested cross-validation could offer a more robust evaluation by reducing bias in model selection and performance estimation.
- **Deep Learning Techniques**:
  Exploring deep learning models, such as neural networks, could potentially capture more complex patterns in the data, especially with large datasets. These models might offer improved accuracy and robustness in predicting depression.

By addressing these aspects, future research can further refine the predictive models and potentially achieve even higher accuracy and reliability in identifying individuals at risk of depression. This project lays a solid foundation for continued exploration and improvement in the field of mental health prediction using machine learning.