ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 5 No. 1 (2025) 334-345
https://publisher.uthm.edu.my/periodicals/index.php/ekst

A Comparative Analysis of Depression in Malaysia and Argentina Using Machine Learning Approach

Siti Hafsah Habeeb Mohamed[1], Sabariah Saharan[1]*,

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: sabaria@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2025.05.01.040

Article Info

Received: 30 December 2024
Accepted: 21 January 2025
Available online: 30 July 2025

Keywords

Depression, Machine Learning, Malaysia, Argentina, CART, RF

Abstract

This study conducts a comparative analysis of depression in Malaysia and Argentina using a machine learning approach. By analysing datasets from the National Health and Morbidity Survey (Malaysia) and the Open Science Framework (Argentina), the study identifies key factors contributing to depression and evaluates the performance of various machine learning models, including Classification and Regression Tree (CART), Random Forest (RF), k-Nearest Neighbour (k-NN), Naïve Bayes (NB), and Support Vector Machine (SVM). The findings reveal socio-economic and demographic variables significantly influencing depression in both countries. CART and RF emerged as the most effective algorithms, where CART achieving the highest accuracy (97.41% for Malaysia and 83.03% for Argentina) and F1-scores (96.13% for Malaysia and 83.22% for Argentina). SMOTE oversampling significantly improved model performance by addressing class imbalance. The results highlight the importance of culturally sensitive predictive models for mental health and provide valuable insights for targeted interventions and policies in Malaysia and Argentina.

1. Introduction

Depression, often termed a silent epidemic, profoundly impacts the lives of countless individuals worldwide. This complex disorder, characterized by persistent sadness, hopelessness, and disinterest in daily activities, affects people of all ages and socioeconomic backgrounds [1]. Its toll on healthcare systems and society underscores the urgent need for effective strategies to address this pressing issue.

Scientific advancements in recent years have highlighted the multifactorial nature of depression, with genetic predispositions, environmental influences, and psychosocial factors playing pivotal roles in its onset and severity [2]. These discoveries emphasize the importance of innovative approaches to better understand and treat this condition.

In Malaysia, mental health issues have led to greater efforts to raise awareness and improve access to care, showing how important mental well-being is for a good quality of life [3]. Despite these efforts, there are still gaps in understanding and managing depression, especially in rural areas where mental health services are harder to access. Cultural stigma also makes it challenging for people to seek help. Malaysia's diverse population of Malay, Chinese, and Indian communities adds complexity, as different cultural beliefs shape how mental health is viewed and addressed. Socio-economic factors, like income differences and the urban-rural divide, further affect the levels and treatment of depression, highlighting the need for targeted, culturally sensitive solutions.

Similarly, mental health challenges are worsened by socioeconomic inequality and poor access to care, especially in rural areas [4]. The country's mix of European and indigenous cultures gives insight into how

depression affects different groups of people. The COVID-19 pandemic made things worse, with higher rates of depression, anxiety, and stress reported during long periods of isolation [5]. Argentina's health data, from the Open Science Framework (OSF), provides an opportunity to create predictive models to better understand and address depression. The availability of this dataset has played a crucial role in analysing the impact of socio-economic factors, cultural diversity, and pandemic-induced stress, offering valuable lessons for improving mental health care and creating more effective, data-driven solutions.

Machine learning-based predictive modelling offers hope in addressing these challenges. By leveraging data-driven approaches, healthcare providers can gain deeper insights into the complex factors contributing to depression [6]. Advanced algorithms can continuously learn and improve from new data, enhancing their accuracy in predicting depressive symptoms and tailoring treatments to individual needs [7]. This ability to personalize care represents a transformative shift in mental health management.

This study seeks to fill critical gaps in understanding and predicting depression in Malaysia and Argentina by focusing on three key areas such as identifying the specific variables contributing to depression in each country, developing advanced machine learning models for early detection, and conducting comparative analyses to understand similarities and differences in mental health outcomes. By examining depression in two distinct socio-cultural and economic contexts, the research aims to enhance culturally sensitive approaches and advance mental health studies in diverse populations.

2. Methodology

2.1 Data

Data for Malaysia is sourced from the National Institute of Health (NIH) via the National Health and Morbidity Survey (NHMS, 2019), comprising 5,149 entries with 9 variables. Table 1 outlines key variables, such as state, gender, age group, and depression status. Similarly, data from Argentina, retrieved from the Open Science Framework (OSF), includes 1,101 entries with 10 variables, including education, mental health history, and depression scores in Table 2. Both datasets are in Excel format and provide crucial insights into the factors influencing depression in each country.

Table 1 Variables in the dataset related to depression in Malaysia

Variables	Description	Data Type
state	State	Categorical
sex	Gender	Categorical
agegp	Age group	Categorical
race	Ethnicity	Categorical
marital	Marital status	Categorical
education	Education level	Categorical
occu_5gp	Occupation group	Categorical
HHincome_gp	Household income group	Categorical
Depression_status	Depression status	Categorical

Table 2 Variables in the dataset related to depression in Malaysia

Variables	Description	Data Type
Education	Education level	Categorical
Sex	Gender	Categorical
Age	Age	Numerical / Continuous
Mental disorder history	Mental health history	Categorical
Suic attempt history	Suicide attempt history	Categorical
Living with somebody	Household members	Categorical
Economic income	Household income	Categorical
Suic risk	Suicide risk	Categorical
Anxiety state	State of anxiety	Categorical
Depression	Depression score	Categorical

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 1 (2025) p. 334-345

2.2 Data Preprocessing

Preprocessing converts raw data into usable information by addressing issues such as noise, redundancy, and imbalance [8]. Certain variables were excluded during preprocessing due to irrelevance to the study. To resolve imbalanced data, oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE) are applied, as they enrich feature diversity and reduce the risk of overfitting [9]. SMOTE specifically handles class imbalance by generating synthetic data for the minority class. Instead of duplicating existing cases, it creates new samples by combining features of a target case with its nearest neighbours. This approach enhances data diversity and ensures a more balanced representation of classes in the dataset [10]. By doing so, SMOTE enables the machine learning algorithms to learn effectively from both majority and minority classes, leading to improved predictive accuracy and fairness. Below are the steps describing how SMOTE works.

SMOTE Steps:
Step 1: Decide the amount of synthetic data (N) needed.
Step 2: Select a minority class data point as a starting point.
Step 3: Find Nearest Neighbours
Step 4: Generate Synthetic Data
Step 5: Repeat until the dataset is balanced.

2.3 Machine Learning Models

2.3.1 Classification and Regression Tree (CART)

CART identifies optimal splits using the Gini Index shown. It is particularly useful for handling categorical and numerical data and provides an interpretable decision-making structure. The Gini Index equation (1) and the data partitioning equation (2) are as follows:

$$Gini\,(D) = \; 1 - \sum_{i=1}^{m} pi^2 \tag{1}$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2}$$

Here, D is the dataset, p is the probability of category i, and D_1, D_2 are binary splits. CART's versatility makes it ideal for diverse datasets [11].

2.3.2 Random Forest (RF)

RF builds multiple decision trees using bootstrapped samples, with final predictions determined by majority voting [12]. This ensemble method enhances generalization and reduces overfitting.

2.3.3 *k*-Nearest Neighbour (*k*-NN)

k-NN classifies data based on proximity to *k* neighbours [13]. he optimal k value is determined through cross-validation to balance bias and variance using equation shown below (3):

$$k = \sqrt{\mathrm{n}} \tag{3}$$

Distance between points x and y is determined using Euclidean distance equation [14]. The equation show in (4):

$$d(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{4}$$

2.3.4 Naïve Bayes (NB)

NB applies Bayes' theorem [15] assuming feature independence. It is particularly efficient for text classification and problems also with categorical variables:

$$P(a|y) = \frac{(P(y|a)(P(a))}{P(y)} \tag{5}$$

where in equation (5) $P(a|y)$ is the posterior probability of outcome a given predictor y.

2.3.5 Support Vector Machine (SVM)

SVM separates classes using a hyperplane, maximizing the margin between data points. This involves two optimization approaches [16] Kernel functions are used to handle non-linearly separable data, making SVM effective for high-dimensional spaces. The steps on SVM works are as follows:

Step 1: Preprocess the dataset.
Step 2: Select a kernel function.
Step 3: Define the margin type.
Step 4: Train the model.
Step 5: Tune the hyperparameters.
Step 6: Evaluate the model.

2.3.6 Model Performance

Examining the performance of a model is crucial once it's developed. Therefore, in this study, the effectiveness of all prediction models, each employing different machine learning algorithms, will be determined using performance indicators such as confusion matrix, accuracy, precision, recall and F1-score [17]. Equations shown in (6), (7), (8), (9) represent the Accuracy, Precision, Recall and F1-score respectively:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

Next, on the variants of F1-score which is the macro and weighted averages equation shown in equation (10) and (11). The macro averages equation:

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{precision_i \times recall_i}{precision_i + recall_i} \tag{10}$$

Followed by the weighted averages:

$$F1_{weighted} = \sum_{i=1}^{N} w_i \times (2 \times \frac{precision_i \times recall_i}{precision_i + recall_i}) \tag{11}$$

where the equation (12) for the w_i:

$$w_i = \frac{TP_i \times FN_i}{\sum_{j=1}^{N}(TP_j + FN_j)} \tag{12}$$

Here, *TP*, *TN*, *FP*, and *FN* represent True Positives, True Negatives, False Positives, and False Negatives, which are the key components used in the calculation of the confusion matrix, recall, precision, and F1-score.

Next, the macro F1-score is calculated by taking the F1-score for each class individually and averaging them. Where in the equation, *N* represents the total number of classes. $precision_i$ refers to the precision for class *i*, and $recall_i$ refers to the recall for class *i*. On the other hand, the weighted F1-score adjusts the average F1-score by giving more importance to larger classes. Where in the equation, TP_i represents the true positives for class *i*, and FN_i represents the false negatives for class *i* [17].

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 1 (2025) p. 334-345

3. Results

3.1 Data

Data preprocessing included variable selection and oversampling. Variable selection reduced Malaysia's variables from 41 to 9 and Argentina's from 13 to 10, targeting relevant factors to avoid overfitting and enhance model performance. Oversampling with SMOTE addressed class imbalance, creating balanced datasets to improve model predictions.

Fig. 1(a) and Fig. 2(a) show the original class distribution in the Malaysia and Argentina datasets, where the majority class had significantly more samples than the minority class. This imbalance could cause the model to favour the majority class, leading to biased results.

After applying SMOTE, Fig. 1(b) and Fig. 2(b) display a balanced class distribution. SMOTE generated additional synthetic data points for the minority class, ensuring an equal representation of both classes. This balancing step helped the model learn from both classes effectively, reducing bias and improving its overall performance.

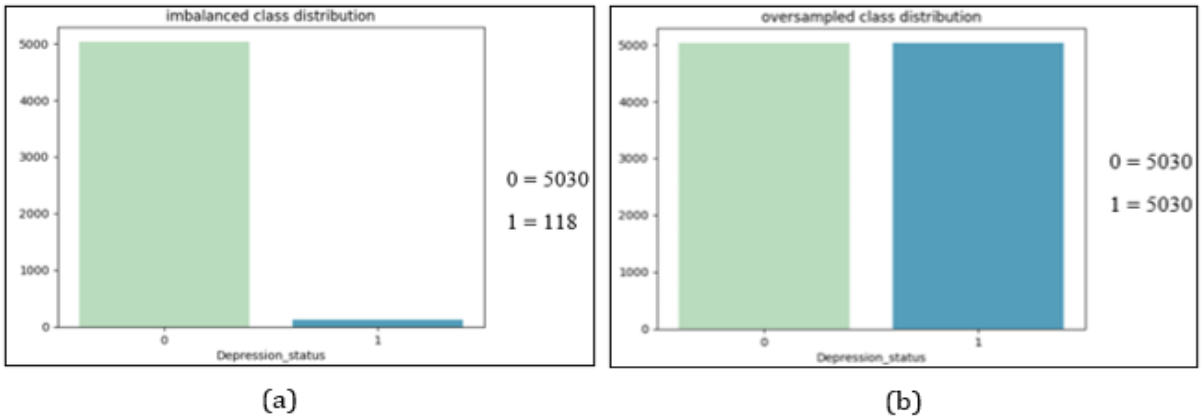

Fig. 1 (a) The imbalanced class distribution for Malaysia dataset (b) The balanced class distribution for Malaysia dataset

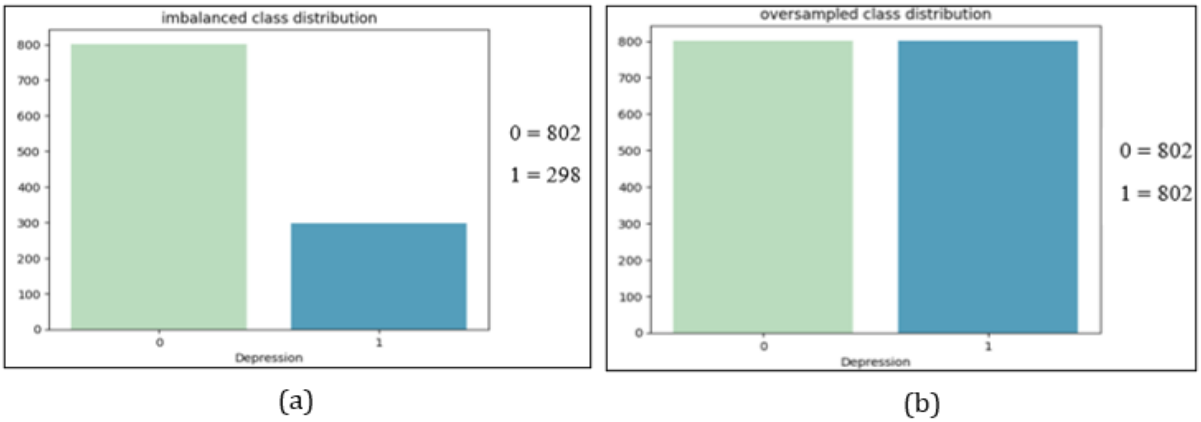

Fig. 2 (a) The imbalanced class distribution for Argentina dataset (b) The balanced class distribution for Argentina dataset

3.2 Factors Affecting Depression in Malaysian Adults (CART)

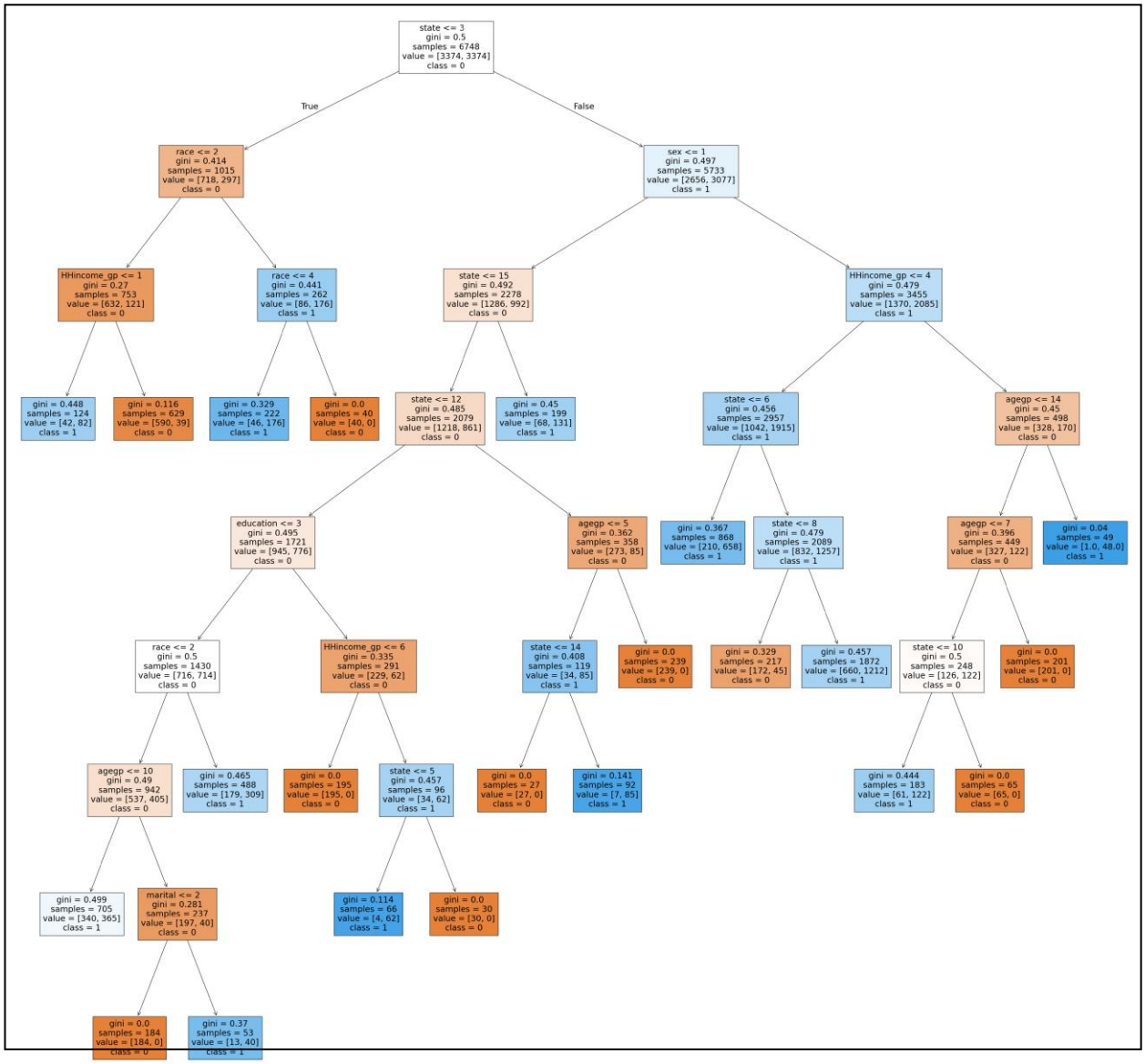

Fig. 3 CART for Malaysia's Data

Fig. 3 shows the Classification and Regression Tree (CART) model applied to the Malaysia dataset which identified several key factors increasing the likelihood of depression among Malaysian adults. The model highlighted geographical locations as significant, with individuals living in states such as Melaka, Negeri Sembilan, Pahang, Penang, Perak, Perlis, Selangor, Terengganu, Sabah, Sarawak, and the Federal Territories (Kuala Lumpur, Labuan, and Putrajaya) are being more likely to experience depression. Gender was also played a crucial factor, with women, particularly those living in the mentioned states and earning a household income below RM4000, exhibiting a higher likelihood of depression. Interestingly, women with lower incomes in other states such as Johor, Kedah, Kelantan, and others also showed an increased depression risk, with this risk being particularly high in Perlis, Selangor, Terengganu, Sabah, Sarawak, and the Federal Territories.

Income level had a notable impact on depression risk with individuals earning above RM4000 generally having a lower chance of depression. However, older individuals aged 65-84, still faced a higher likelihood of depression despite their higher income. Age also played as a contributing factor, with younger individuals (15-69 years old) showing lower depression risk if they had higher incomes, although those in certain states still had significant risk. The model also revealed gender-specific patterns, with men in most states having a lower chance of depression, except for those in Putrajaya. However, young men aged 15-24, especially those in the Federal Territories, exhibited a higher risk.

Education level, particularly low education (no formal education or only primary or secondary education), was associated with higher depression risks, especially in individuals from states like Johor, Kedah, Kelantan, and others, and those identifying as Indian, or coming from minority groups in Sabah and Sarawak. Moreover,

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 1 (2025) p. 334-345

Malay and Chinese individuals aged 15-49 were more likely to experience depression, along with older adults aged 50-84 who were widowed. Finally, specific combinations of state, race, and income were strongly associated with high depression risk, such as Indian individuals in Johor, Kedah, and Kelantan with low household incomes. These insights can guide targeted mental health interventions aimed at groups at higher risk of depression.

3.3 Factors Affecting Depression Among Argentine Adults Using CART

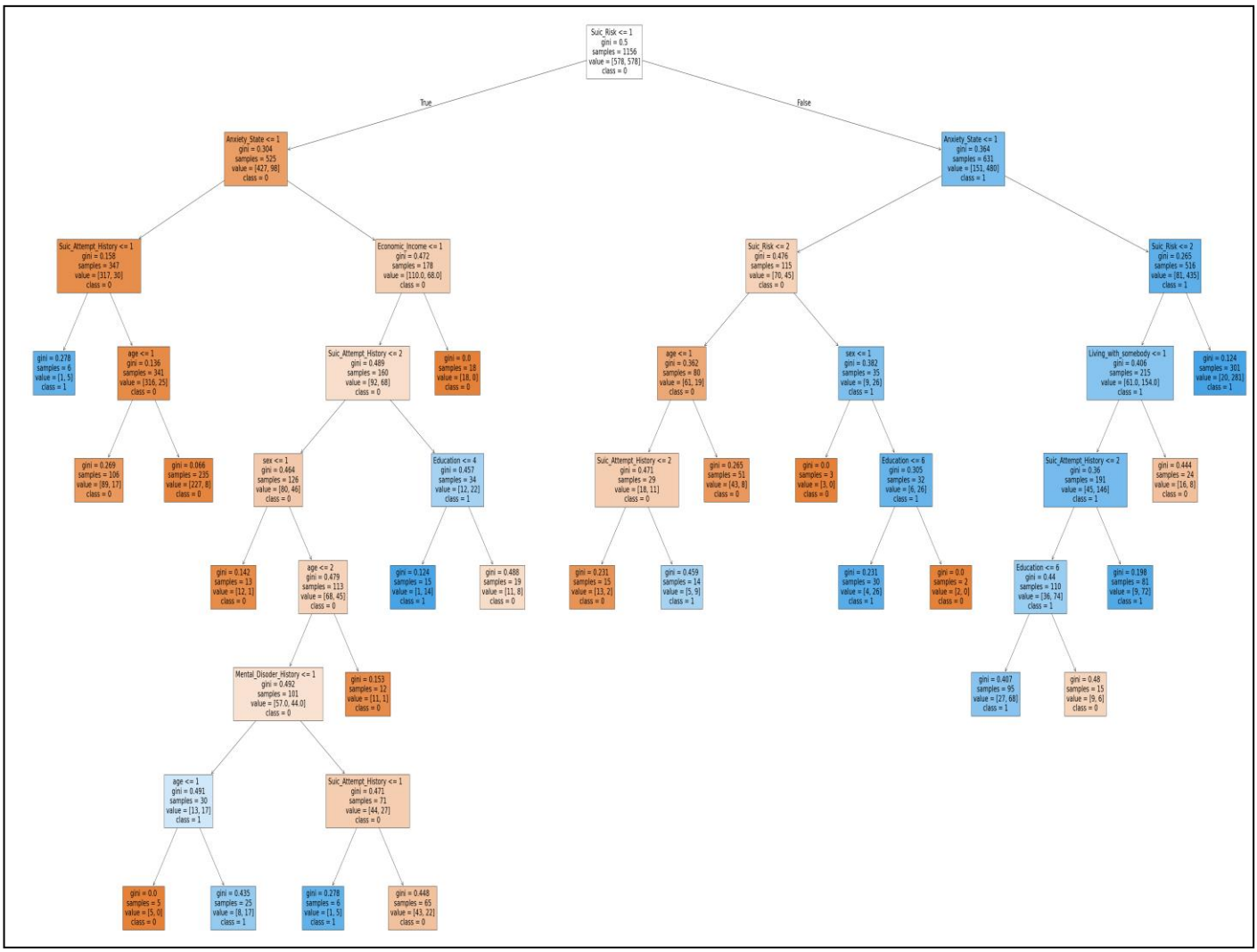

Fig 4. CART for Argentina's Data

The CART model in Fig. 4 applied to the Argentina dataset identified several critical factors that contribute to depression risk among Argentine adults, focusing on anxiety levels, suicide risk, age, education, and economic status. The analysis found that individuals with moderate to high levels of anxiety and those with a low to moderate suicide risk, particularly those who experience suicidal ideation, are more likely to experience depression. Furthermore, depression risk is significantly elevated in individuals who have a high suicide risk and a history of suicide attempts, even when they live with others, suggesting that the presence of a support system does not always mitigate vulnerability to depression.

Educational background was also a significant factor, with individuals who had only a high school education, combined with high suicide risk and previous suicide attempts, being at greater risk of depression. This risk was especially pronounced among females who had low anxiety levels but high suicide risk and a history of suicide attempts. Age also played a role, with individuals aged 18-24, who had low to moderate suicide risk and experienced suicidal ideation, showing an increased likelihood of depression. The model emphasized that prior suicide attempts were a strong indicator of depression, even in cases where the current suicide risk was low. For adults aged 25-44, the combination of stable income and a history of mental health issues suggested a higher risk of depression, although even individuals without a mental health history but with a suicide attempt were at an increased risk. Overall, the CART analysis for Argentina identified a combination of mental health history,

anxiety, age, education, and economic stability as crucial factors that determine depression risk, highlighting the need for targeted mental health programs that address these factors to reduce depression in the population.

3.4 Confusion Matrices for Malaysia Dataset

Table 3 Confusion Matrix of CART algorithms for Malaysia dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	1375	130
Negative (0)	30	10

Table 4 Confusion Matrix of RF algorithms for Malaysia dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	1410	95
Negative (0)	35	5

Table 5 Confusion Matrix of k-NN algorithms for Malaysia dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	1217	288
Negative (0)	29	11

Table 6 Confusion Matrix of NB algorithms for Malaysia dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	1217	288
Negative (0)	29	11

Table 7 Confusion Matrix of SVM algorithms for Malaysia dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	1217	288
Negative (0)	29	11

Table 8 Compilation on Confusion Matrix for Malaysia dataset

Algorithm	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
CART	1375	10	130	30
RF	1410	5	95	35
k-NN	1217	11	288	29
NB	1217	11	288	29
SVM	1217	11	288	29

For the Malaysia dataset as shown in table 8, CART exhibited the highest true positive rate, with 1375 TP, but struggled with lower true negative values. RF balanced TP (1410) and TN slightly better than CART. However, *k*-NN, NB, and SVM had similar performances, with lower overall TP and TN counts, making them less effective for this dataset.

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 1 (2025) p. 334-345

3.5 Confusion Matrices for Argentina Dataset

Table 9 Confusion Matrix of CART algorithms for Argentina dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	191	33
Negative (0)	23	83

Table 10 Confusion Matrix of RF algorithms for Argentina dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	195	29
Negative (0)	31	75

Table 11 Confusion Matrix of k-NN algorithms for Argentina dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	187	37
Negative (0)	26	80

Table 12 Confusion Matrix of NB algorithms for Argentina dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	187	37
Negative (0)	26	80

Table 13 Confusion Matrix of SVM algorithms for Argentina dataset

	Actual Values	
Predictive Values	Positive (1)	Negative (0)
Positive (1)	187	37
Negative (0)	26	80

Table 14 Compilation on Confusion Matrix for Argentina dataset

Algorithm	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
CART	191	83	33	23
RF	195	75	29	31
k-NN	187	80	37	26
NB	187	80	37	26
SVM	187	80	37	26

In the Argentina dataset shown in Table 14, CART again led, achieving a strong TP (191) and TN (83). RF closely followed with higher TP (195) but slightly lower TN (75). *k*-NN, NB, and SVM showed comparable performances, being slightly more effective in non-depression cases but weaker in depression detection.

3.6 Performance Matrices for Malaysia and Argentina

Table 15 Performance Matrix for Malaysia dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CART	97.41	94.89	97.41	96.13
RF	91.59	95.18	91.59	93.30
k-NN	79.48	95.24	79.48	86.35
NB	79.48	95.24	79.48	86.35
SVM	79.48	95.24	79.48	86.35

Table 16 Confusion Matrix for Argentina dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CART	83.03	83.57	83.03	83.22
RF	81.82	81.73	81.82	81.77
k-NN	80.91	81.56	80.91	81.14
NB	80.91	81.56	80.91	81.14
SVM	80.91	81.56	80.91	81.14

Based on Tables 15 and 16 shown, CART consistently outperformed other algorithms in both datasets. For the Malaysia dataset, it achieved the highest accuracy (97.41%) and F1-Score (96.13%), with RF following closely. Similarly, in the Argentina dataset, CART excelled with an accuracy of 83.03% and F1-Score of 83.22%. RF showed balanced results, but CART maintained superiority.

3.7 F1-Score Variants Summary

Table 17 F1-Score Variants for Malaysia dataset

Algorithm	Macro Average F1-Score (%)	Weighted Average F1-Score (%)
CART	49	96
RF	53	94
k-NN	47	86
NB	47	86
SVM	47	86

Table 18 F1-Score Variants for Argentina dataset

Algorithm	Macro Average F1-Score (%)	Weighted Average F1-Score (%)
CART	81	83
RF	78	81
k-NN	79	81
NB	79	81
SVM	79	81

Macro and weighted F1-scores demonstrated CART's strength in balanced evaluations as shown in Table 17 and 18. In Malaysia, RF had the highest macro average (53%), while CART dominated the weighted F1-score (96%). For Argentina, CART led in both macro average (81%) and weighted F1-score (83%), indicating its reliability in handling both balanced and imbalanced data conditions.

To sum up, according to the F1 score and its variants, CART outperformed other models, demonstrating high accuracy and ease of interpretation. This suggests decision tree models are effective in identifying key depression risk factors. These findings can help improve mental health policies by guiding resource allocation and targeted intervention programs. Governments and health organizations can use this information to support high-risk groups and enhance early detection efforts through machine learning tools.

4. Conclusion

This study evaluated the effectiveness of machine learning algorithms CART, RF, k-NN, NB, and SVM in predicting depression risk in Malaysia and Argentina. Among these, CART stood out as the best-performing model in both datasets, leading in metrics such as accuracy, recall, and F1-score. Additionally, CART demonstrated the ability to visualize the final decision tree, providing clear insights into the key factors influencing depression risk. This makes CART a reliable and interpretable tool for identifying depression cases

in diverse populations. RF also performed well but did not perform as well as CART, while k-NN, NB, and SVM showed lower effectiveness, particularly in detecting depression cases.

The use of SMOTE during preprocessing was critical in addressing class imbalances, ensuring fair representation of minority and majority classes. This contributed to the balanced results reflected in the macro and weighted average F1-scores, highlighting CART's ability to treat all classes equitably while maintaining strong overall accuracy. SMOTE's role underscores the importance of preprocessing techniques like oversampling in improving model performance and generating meaningful insights.

To maximize the impact of these findings, healthcare providers should integrate machine learning models like CART into early screening programs to improve depression detection and intervention. Policymakers should allocate resources to implement AI-driven mental health assessments in public health initiatives. Future research should explore datasets from other regions to understand how depression risk factors vary across different populations. Additionally, advanced machine learning models like Neural Networks, XGBoost, and LightGBM should be investigated to further enhance predictive power and support mental health analysis.

Acknowledgement

The author thanks Universiti Tun Hussein Onn Malaysia for supporting this study.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Siti Hafsah Habeeb Mohamed, Sabariah Saharan;* ***data collection****: Siti Hafsah Habeeb Mohamed, Sabariah Saharan;* ***analysis and interpretation of results****: Siti Hafsah Habeeb Mohamed, Sabariah Saharan;* ***draft manuscript preparation****: Siti Hafsah Habeeb Mohamed, Sabariah Saharan. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] World Health Organization. (2023). *Depressive disorder (depression)*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/depression

[2] Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & Griffiths, S. L. (2024). The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58–90. https://doi.org/10.1002/wps.21160

[3] Ibrahim, N., Mohd Safien, A., Siau, C. S., & Shahar, S. (2020). The Effectiveness of a Depression Literacy Program on Stigma and Mental Help-Seeking Among Adolescents in Malaysia: A Control Group Study With 3-Month Follow-Up. *The Journal of Health Care Organisation, Provision and Financing*, *57*.

[4] Mitelman, M., Chirazi, A., Schmollgruber, A., & Leiderman, E. A. (2023). Discrimination and social stigma against people with mental illnesses in Argentina. *International Journal of Social Psychiatry, 69*(2), 334-341.

[5] Badellino, H., Gobbo, M. E., Torres, E., Aschieri, M. E., Biotti, M., Alvarez, V., Gigante, C., & Cachiarelli, M. (2021). 'It's the economy, stupid': Lessons of a longitudinal study of depression in Argentina. *Int J Soc Psychiatry*, 384-391.

[6] Chahar, R., Dubey, A. K., & Narang, S. K. (2021). A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. In *International Journal of Advanced Technology and Engineering Exploration* (Vol. 8, Issue 83, pp. 1279–1314). Accent Social and Welfare Society. https://doi.org/10.19101/IJATEE.2021.874198

[7] Lee, C., & Kim, H. (2022). Machine learning-based predictive modeling of depression in hypertensive populations. *PLoS ONE*, *17*(7 July). https://doi.org/10.1371/journal.pone.0272330

[8] Adari, G. K., Raja, M., & Vijaya, P. (2023). Machine learning in genomics: identification and modeling of anticancer peptides. *In Data Science for Genomics* (pp. 25-68). Academic Press.

[9] Ali, H., Salleh, M. N. M., Hussain, K., Ahmad, A., Ullah, A., Muhammad, A., ... & Khan, M. (2019). A review on data preprocessing methods for class imbalance problem. *International Journal of Engineering & Technology*, *8*(3), 390-397.

[10] Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, *10*, 47643-47660.

[11] Gomes, C. M., Lemos, G. C., & Jelihovschi, E. G. (2020). Comparing the predictive power of the CART and CTREE algorithms. *Avaliação Psicológica, 19*(1), 87-96.

[12] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1
[13] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071. https://doi.org/10.1016/j.dajour.2022.100071
[14] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing, 8*, 341-357.
[15] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x
[16] Muzzammel, R., & Raza, A. (2020). A support vector machine learning-based protection technique for MT-HVDC systems. *Energies*, *13*(24). https://doi.org/10.3390/en13246668
[17] Cao, C., Chicco, D., & Hoffman, M. M. (2020). The MCC-F1 curve: a performance evaluation technique for binary classification.