VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

Subject

CS313 - Data Mining and Application

FINAL REPORT

Understanding and Predicting Depression to Enhance Mental Health Interventions

Advisor: Vo Nguyen Le Duy
Class name: CS313.P11
Group 8:

1. Le Tran Bao Loi - 21522295
2. Pham Van Hung - 21522124
3. Do Ba Huy - 21522137
4. Tran Vi Khang - 21522201
5. Le Dang Khoa - 21522222
6. Do Phuc Kien - 21522243
7. Nguyen Duc Lap - 21522295

HO CHI MINH CITY, Dec 2024

Contents

1 Introduction

Mental health is becoming an increasingly significant global concern. In Vietnam alone, over 40,000 suicides due to depression are recorded annually. However, this number is just the tip of the iceberg, as the actual number of people suffering from depression is much higher and often goes undetected until it is too late. Therefore, we consider it crucial to identify and assist those at risk of depression in a timely manner. We have decided to develop a model to predict whether a person is at risk of depression based on certain biological characteristics and lifestyle habits.

The synthetic dataset we utilized originates from a mental health survey, offering an opportunity to explore these factors and develop predictive models that can aid in early detection.

From these data containing information about survey participants, we will identify the key features and build a predictive model based on those features.

2 Problem Formulation

2.1 Problem Statement

How can us leverage data analytics to identify key factors associated with depression and predict the likelihood of depression in individuals? Specifically:

- Which variables have the strongest correlation with the risk of depression?
- Can a machine learning model accurately predict the likelihood of depression based on survey data?
- How can these predictions be utilized to inform policy-making or design targeted mental health programs effectively?

2.2 Objective

Develop a machine learning-based solution that aims to:

- Identify the key factors contributing to depression from a mental health survey dataset.
- Accurately predict the likelihood of depression for new unseen data.
- Provide actionable insights that can help organizations, such as mental health organizations, policymakers, and healthcare providers, design tailored interventions to address depression effectively.

2.3 Significance and Practical Implications

Addressing this problem can lead to several significant outcomes:

1. **Optimizing resource allocation:** Enable organizations to identify high-risk groups and prioritize the distribution of resources to areas or populations most in need.

2. **Raising community awareness:** Facilitate the creation of awareness campaigns that target specific risk factors associated with depression.

3. **Improving program effectiveness:** Provide data-driven insights to monitor and evaluate the impact of mental health programs, ensuring they are evidence-based and effectively address the needs of the population.

3 Proposed Approaches

The proposed approaches involve comprehensive steps to handle data and build predictive models. First, data preprocessing was conducted, including handling missing values, dropping unnecessary columns (e.g., id and Name), converting data types, and encoding categorical data using LabelEncoder. Next, exploratory data analysis (EDA) was performed to analyze relationships between features and the target variable (Depression), supported by visualizations such as distribution plots, heatmaps, and bar charts. Outlier detection was carried out using the Interquartile Range (IQR) method [2], followed by data normalization using StandardScaler.

Various machine learning models were explored, including Decision Tree, Random Forest [5] and LightGBM [4]. These models were trained and evaluated using metrics such as accuracy, F1-Score, and AUC-ROC. Additionally, a Multi-Layer Perceptron (MLP) [6] neural network was designed using TensorFlow/Keras, employing ReLU [1] and sigmoid activation functions, with dropout layers added to prevent overfitting.

Finally, models were thoroughly evaluated through classification reports, confusion matrices, and ROC curves [3], enabling performance comparison to identify the best-performing approach for predicting depression.

4 Experiment

4.1 Data Analysis Stage

4.1.1 Data Cleaning:

There are many missing values in this dataset. If the data is not cleaned beforehand, the model may perform inaccurately, leading to results that deviate from expectations.

Feature	Percentage of Missing Values
id	0.000000
Name	0.000000
Gender	0.000000
Age	0.000000
City	0.000000
Working Professional or Student	0.000000
Profession	26.034115
Academic Pressure	80.172708
Work Pressure	19.842217
CGPA	80.171997
Study Satisfaction	80.172708
Job Satisfaction	19.836532
Sleep Duration	0.000000
Dietary Habits	0.002843
Degree	0.001421
Have you ever had suicidal thoughts ?	0.000000
Work/Study Hours	0.000000
Financial Stress	0.002843
Family History of Mental Illness	0.000000
Depression	0.000000

Table 1: Percentage of Missing Values in Each Variable

To prevent this, we clean the data by removing columns with a high proportion of missing values (where the missing value rate exceeds 70%) and imputing the missing values for columns with fewer missing values. For numerical columns, we choose to impute missing values using the median because it is less affected by outliers compared to the mean. For categorical columns, we choose to impute missing values using the mode (the most frequent value), which helps preserve the distribution characteristics of the data.

4.1.2 Exploding Data Analysis:

- Target: The rate of non-depression (82%) is significantly higher compared to depression (18%). A significant imbalance in the target ratio like this often causes the model to be biased toward the majority class, making it challenging to learn the differences between individuals with and without depression. This could lead to an increase in False Negatives, which would seriously impact the goal of identifying those at risk of depression.

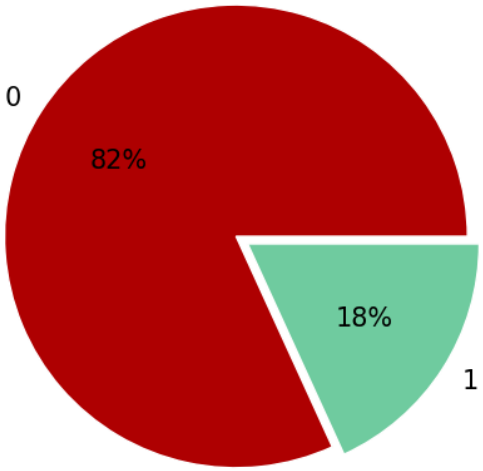
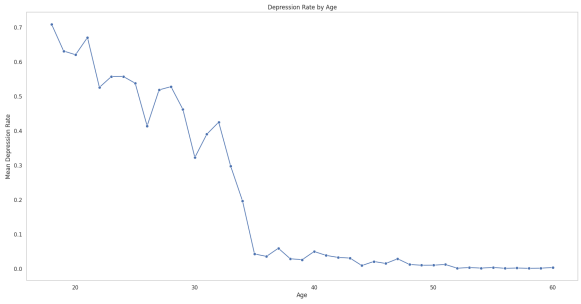
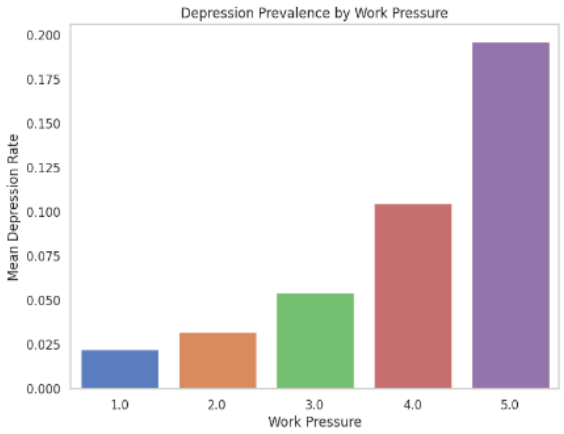


Figure 1: Target Variable Distribution

- Numerical:



Depression rate by Age

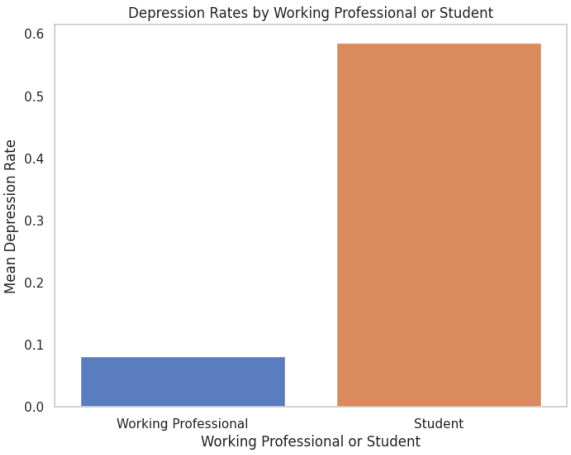


Depression rate by Work Pressure

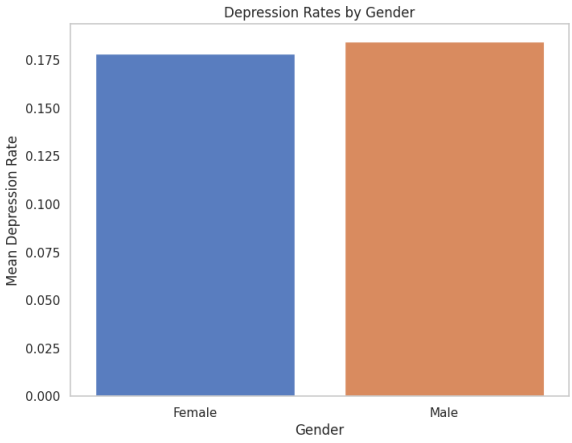
Figure 2: Depression rates by Age and Work Pressure

Based on the analysis results, certain features clearly differentiate the number of people with depression. For instance, individuals aged 30 and below constitute the majority of those with depression, and those experiencing greater financial stress are more likely to be depressed. However, some features do not exhibit such linear differentiation. For example, individuals with moderate work pressure and job satisfaction account for the highest proportion of depressed individuals. With this data, we can focus more on groups with higher depression rates to monitor signs of depression effectively.

- Categorical:



Depression rate by Worker/Student



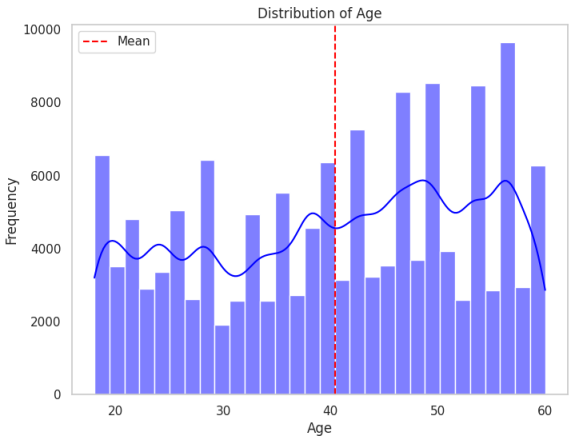
Depression rate impact by Gender

Figure 3: Depression rates by Worker/Student and Gender

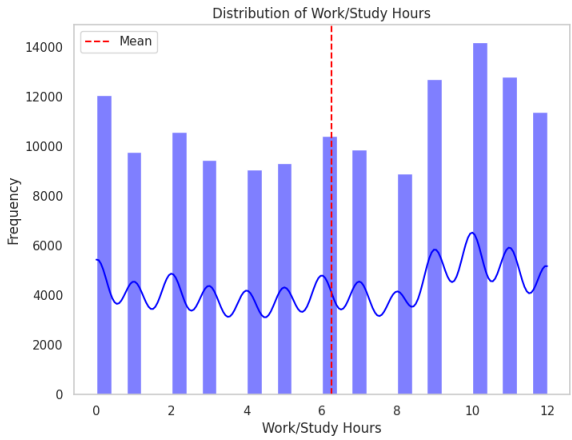
Similar to numerical features, categorical features also help distinguish groups at risk of depression. For example, students account for a higher proportion of depressed individuals compared to those already working. Additionally, certain factors, such as gender, appear to have minimal impact on depression rates.

4.1.3 Outliers Analysis:

The data in the numerical columns is very evenly distributed, with no outliers detected in any of the columns, even those containing missing values that were filled using the median. This demonstrates that the median imputation method effectively minimizes the presence of outliers.



Distribution of Age



Distribution of Work/Study Hours

Figure 4: Distribution of Age and Work/Study Hours

4.2 Predictive Model Stage

In this stage, we focused on constructing, training, and evaluating a variety of machine learning and deep learning models to effectively predict the target variable. Additionally,

feature importance analysis was conducted to determine the contributions of individual features to the model predictions.

4.2.1 Build and Train model

To construct and train the predictive models, the following steps were undertaken:

1. Data Splitting: The dataset was divided into training and testing subsets using an 80:20 split. This approach ensures that the models are evaluated on unseen data, enabling robust generalization.

2. A diverse set of machine learning models was chosen to ensure comprehensive evaluation, including:

   - **Decision Tree Classifier**: A simple, interpretable model prone to overfitting. To ensure reproducibility, the random_state parameter was set to 42 during training.
   - **Random Forest Classifier**: An ensemble model combining multiple decision trees to reduce overfitting and enhance accuracy. It was configured with 100 estimators (n_estimators=100) and random_state is 42 for consistent results.
   - **Gradient Boosting Models**: Advanced ensemble methods were employed for high-performance classification tasks:
     - LightGBM: Leveraging boosting_type is gbdt for gradient boosting, we optimized performance and mitigated overfitting using bagging_fraction and bagging_freq is 0.9 and 0.05, respectively. The model was trained with a learning_rate is 5x10e-2 for balanced learning dynamics.

3. To compare with the machine learning models, we implemented a feedforward neural network (Multilayer Perceptron, MLP) using TensorFlow for binary classification. The architecture included:

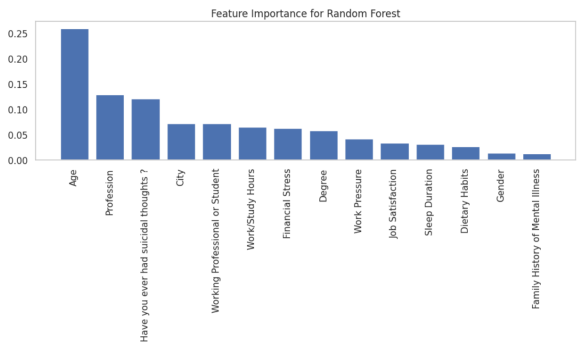
   - Input Layer: Matching the dimensionality of the scaled training data.
   - Hidden Layers: Two layers: the first with 64 neurons and the second with 32 neurons, both utilizing the ReLU activation function. Dropout layers with a rate of 0.3 were added after each hidden layer to mitigate overfitting.
   - Output Layer: A single neuron with a sigmoid activation function for binary output.

   The model was compiled using the Adam optimizer (learning_rate=0.001) with a binary cross-entropy loss function and accuracy as the evaluation metric. Training was performed on the scaled dataset (X_train_scaled), with 20% of the data reserved for validation. The model was trained over 30 epochs with a batch size of 32, providing a balance between computational efficiency and convergence. Dropout layers and validation monitoring ensured robust performance while reducing overfitting risks.

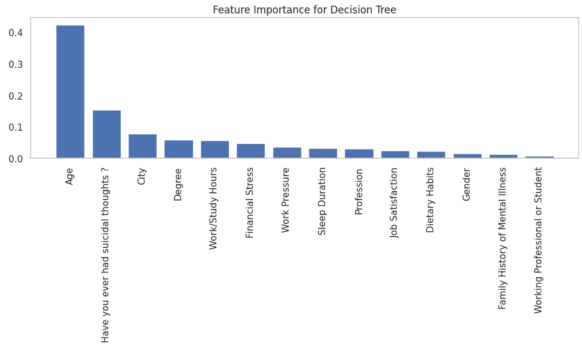
By combining traditional machine learning models and a deep learning approach, this stage provides a thorough comparison and comprehensive evaluation of predictive performance, enabling us to identify the most effective method for the task.

4.2.2 Feature Importances

In this problem, we aim to compare and analyze the feature importance derived from three machine learning algorithms: Random Forest, Decision Tree. Each algorithm provides insights into the relative significance of various features in predicting mental health-related outcomes. The feature importance visualizations are labeled as Figure 5, corresponding to the Random Forest, Decision Tree models, respectively.



Feature Importance for Random Forest



Feature Importance for Decision Tree

Figure 5: Feature importance comparison between Random Forest and Decision Tree

Based on the feature importance charts for both the Random Forest and Decision Tree models, Age stands out as the most influential feature, significantly surpassing other factors. This underscores the critical role of age in predicting the target variable. Additional features, such as Profession and Have you ever had suicidal thoughts?, also demonstrate notable importance, although their rankings vary between the two models.

Based on the feature importance charts for both the Random Forest and Decision Tree models, Age stands out as the most influential feature, significantly surpassing other factors. This underscores the critical role of age in predicting the target variable. Additional features, such as Profession and Have you ever had suicidal thoughts?, also demonstrate notable importance, although their rankings vary between the two models.

Notable differences in feature importance include City and Degree, which are emphasized more in the Decision Tree but receive less attention in the Random Forest model. On the other hand, features such as Gender and Family History of Mental Illness consistently rank as the least important in both models, suggesting they hold minimal predictive value.

In summary, Age emerges as the most critical feature across both models and should be prioritized in predictive analysis. Random Forest, with its more balanced evaluation of feature importance, offers a more robust approach for feature selection. To simplify the models and improve efficiency, features with minimal importance - such as Gender and Family History of

Mental Illness - could be considered for removal. .

4.3 Evaluation Stage

For this problem, we use three metrics for evaluation: Accuracy, F1-Score, and AUC-ROC.

- Accuracy measures the overall correctness of the model by calculating the ratio of correct predictions (True Positives and True Negatives) to the total number of predictions. It is easy to understand but may be misleading in imbalanced datasets.

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

- F1-Score is the harmonic mean of Precision and Recall, balancing the trade-off between detecting positive cases and avoiding false positives. It is particularly useful in imbalanced classification tasks.

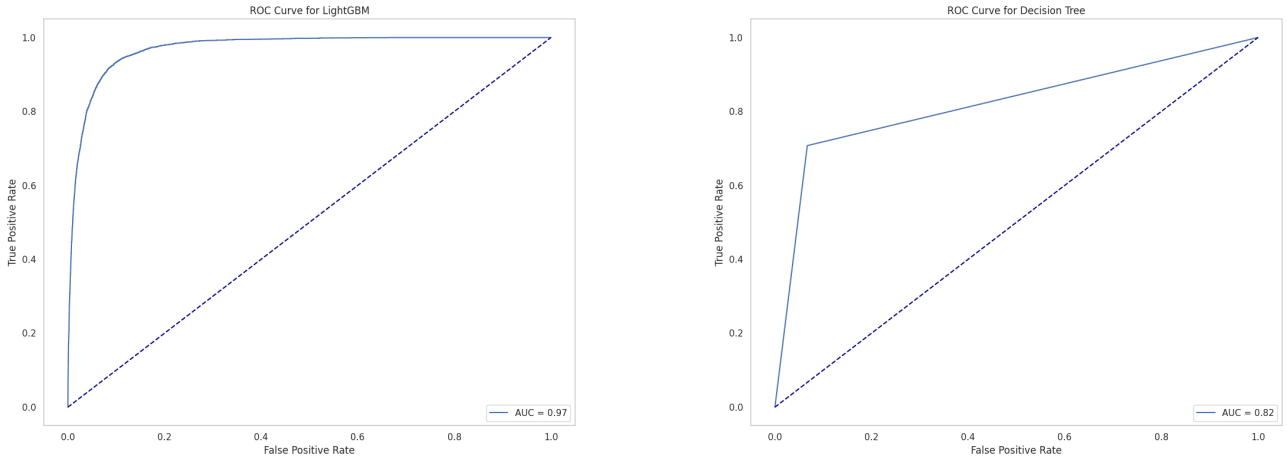
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- AUC-ROC measures the model's ability to distinguish between classes by evaluating the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds. It is unaffected by class imbalance and reflects the model's discriminatory power.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

ROC curves give a more balanced view of a model's ability to distinguish between classes. The Area Under the ROC Curve (AUC) provides a single metric that reflects the model's overall discrimination power, making it ideal for assessing performance in such scenarios.



ROC Curve for LightGBM

ROC Curve for Decision Tree

Figure 6: ROC Curve for LightGBM and Decision Tree

The AUC values close to 1 in the models indicates that the data have been well-cleaned and important features have been effectively selected, allowing the model to overcome the challenges of imbalanced classification.

Specifically, the values of each metric when evaluating the models are as follows:

Model	Accuracy	F1-Score	AUC-ROC
Decision Tree	0.887846	0.697817	0.817707
Random Forest	0.927257	0.796460	0.967375
LightGBM	0.930384	0.808148	0.970895
MLP	0.929602	0.801932	0.874492

Table 2: Model Performance Metrics

4.4 Demo

We are demonstrating this project on the Streamlit platform. Streamlit provides an easy-to-use interface that allows us to showcase the functionality and features of our project in an interactive and engaging way. By leveraging Streamlit, we can quickly build and present data-driven applications with minimal coding, making it an ideal choice for this demonstration.

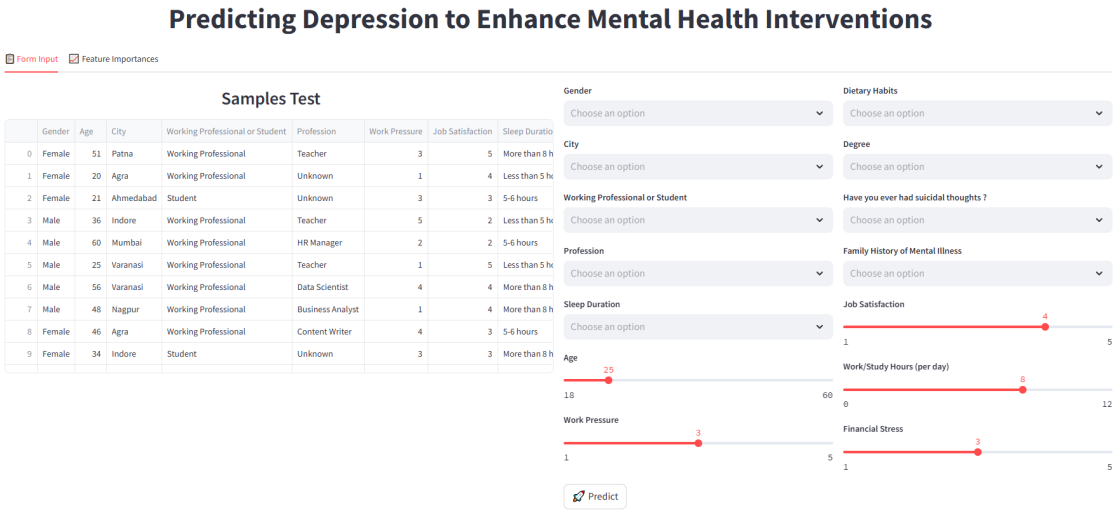


Figure 7: User interface screen

5 Conclusion and Development direction

5.1 Conclusion

Due to careful data preprocessing and the identification of Feature Importances, even with a large discrepancy in the target ratio and the use of simple models for training, it is still possible to achieve very good performance in detecting individuals with depression.

5.2 Development direction

- Data Expansion: Collect more data from diverse sources to improve the accuracy and generalizability of the predictive model.
- In-depth Analysis: Further investigate specific factors that strongly influence depression, such as social, economic, and environmental factors.
- Interdisciplinary Collaboration: Strengthen collaboration between researchers, healthcare providers, and government organizations to develop more comprehensive solutions.

6 Reference

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

[2] Ch. Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, and Ashish Ghosh. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6:100164, 2023.

[3] Tilmann Gneiting and Peter Vogel. Receiver operating characteristic (roc) curves, 2018.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[5] Gilles Louppe. Understanding random forests: From theory to practice, 2015.

[6] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.