University of Jeddah
College of Computer Science and Engineering
Department of Computer Science and Artificial Intelligence
CCAI-312 | Pattern Recognition

Predicting suicide in postpartum depression patients

Layan Alsaud | 2110946, Maryam Aldahri | 2110643, Mariyyah Samarin | 2111188

CCAI-312 | Pattern Recognition Course project

Dr. Safa Alsafari

Jun. 4, 2023

| Introduction

The healthcare sector has developed rapidly in recent years, owing partly to advances in artificial intelligence. Machine learning has proven to be an efficient diagnostic tool for many medical disorders. However, it is becoming more critical to utilize machine learning algorithms in mental health, such as postpartum depression, which affects many new moms. It can result in depression, anxiety, and in severe cases, even suicidal thoughts or acts. Postpartum depression must be identified and treated to prevent terrible outcomes. We attempt to analyze a few parameters, such as age, patient feelings, and problems with attachment to the newborn, to create accurate models that can pinpoint which patients are most in danger of suicide.

| Problem Description

Postpartum depression is a type of clinical depression that occurs after childbirth. It occurs in about 6.5% to 20% of women in the first year after giving birth due to the rapid drop in hormones after delivery there are a lot of risk factors external to the body that can lead to PPD, and there are still some who deny its existence or minimize its severity. These individuals may argue that it is simply a normal thing or that it is an excuse for women, it is important to discover the symptoms early and seek help from a healthcare professional to prevent the worst outcome

| Data Description

Input:

- Age: age of the patient (range=35-40,40-45, other).
- Feeling sad or Tearful: feelings of the patient (yes, no, other).
- Irritable towards baby & partner: patient feeling impatient (yes, no, other).
- Trouble sleeping at night: patient problem to sleep (Two or more days a week, yes, other).
- Problems concentrating or deciding: patient problem in focusing or making a decision (no, often, other).
- Overeating or loss of appetite: patient problem with eating (no, yes, other).
- Feeling anxious: feeling of the patient (yes, no).
- The feeling of guilt: feeling of the patient (no, maybe, other).
- Problems of bonding with baby: patient has a problem with Attachment (no, sometimes, other).

Output (Target attribute): Suicide attempt: patient problem with the will to live (no, yes).

| Decision Tree Model

A Decision Tree is a supervised learning algorithm that can be used for classification or regression and follows a tree-like format, where a node represents a distinct attribute/feature, a branch represents a decision rule, and leaf nodes represent a label. Decision trees have several benefits, including their ease of explanation through visualization and their ability to handle both numerical and categorical data without the need for data preparation such as normalization. However, decision trees also have some weaknesses, such as the potential to overfit the data by creating overly complex trees that learn from noise, and their sensitivity to small changes in the data which can have a large effect.

| Support Vector machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression tasks. The basic idea of the SVM algorithm is to find the maximum distance, called margins, between the closest points in different classes which are called support vectors. This is achieved by drawing a hyperplane that separates data into the appropriate classes. While SVM was originally designed for binary classification problems, it can now be used for multiclass classification problems as well. The SVM can handle linear, non-linear, and radial basis functions (RBF) by using a kernel function that maps data into higher dimensional space.

Some benefits of the SVM classification problem :

-Effective for high-dimensional data -Robust to noise

-Flexibility -Good generalization

Some weaknesses of the SVM classification problem :

-Sensitivity to the choice of kernel -Limited to binary classification

- Computationally intensive

The SVM is a powerful model that has been used in a wide range of applications, such as text classification, image classification, medical diagnosis, speech recognition, and social network analysis. To achieve optimal performance, the SVM algorithm requires well-tuning and evaluation.

| Method

The postpartum depression data set was collected through a questionnaire about postpartum depression for new mothers by a medical hospital and published in Kaggle. However, before constructing a supervised learning classifier, we need to understand the data comprehensively, detect any patterns and correlations among variables and pre-process data to handle any missing values all these steps are integral components of exploratory data analysis (EDA). The next point to consider is selecting an appropriate classifier for the problem. Our objective is binary classification, to predict whether the woman is at risk of suicide or not, so we have chosen to use a decision tree classifier, as part of EDA, and before the training phase, we need to transform data from categorical to numerical format using techniques such as one-hot encoding, as a decision tree can handle only numerical values, in addition to split data into train and test data to evaluate the model performance. Moreover, after training the classifier using random hyperparameters, we then try to fine-tune them by using RandomizedSearchCV to identify the best hyperparameters for optimal model performance. We use different evaluation matrices to measure model performance as F1 score, precision, and recall. Finally,we build an advanced model to compare the performance of the decision tree against it.

| Experiment and results

In this problem, the confusion matrix is used to evaluate the performance of the Decision Tree Classifier and SVM models. The confusion matrix has four possible outcomes in the binary classification problem (as you can see in Figure 1) true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Also, the confusion matrix helps to calculate precision, recall, and F1-score. Finally, the performance of the models will be high when the FP and FN have a minimum value in the problem.

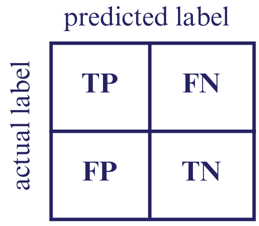

Figure 1: confusion matrix

The values of the confusion matrix for the Decision Tree Classifier are shown in Figure 2 before tuning the hyperparameters, while Figure 3 displays the values after tuning the hyperparameters. We use RandomizedSearchCV to tune hyperparameters. It is a computationally expensive process and sometimes misses the optimal hyperparameters hence it does not cover all possible combinations of hyperparameters. But, it can improve the performance of DT and reduce the risk of overfitting as is shown in our case. Moreover, These values are indicative of the model's classification performance, our model predicts if the patient tends to suicide or not, so we need to decrease FN values. In other words, if the model predicts that the patient is mentally healthy and mismatches the reality it will cost the patient's life. In addition, Figure 4 presents the values for the SVM model which are TP=148, TN=226, FN=2, and FP=6. These values indicate that the SVM model is capable to classify the different classes with a high degree of performance and lower FN compare to the Decision Tree Classifier.

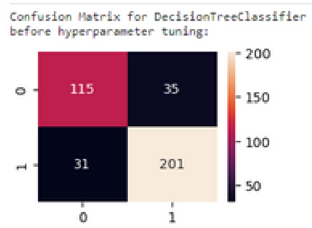

Figure 2: CM_DT1

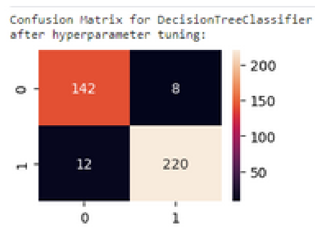

Figure 3: CM_DT2

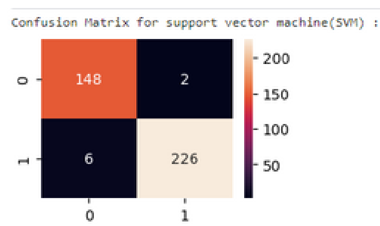

Figure 4: CM_SVM

Finally , We can conclude from Figure 5 that there is no significant difference in values between the train and test data, the model is not overfitting. Also , the errors for both values are low so the model is not underfitting . we can conclude that the model can generalize unseen data well.

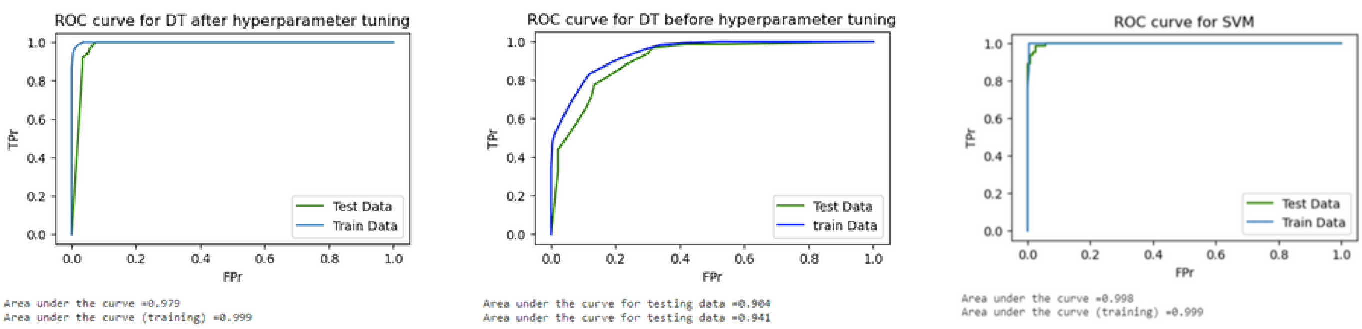

Figure 5: ROC & AUC for classifiers

| Discussion

When comparing the performance of the Support Vector Machine (SVM) and the Decision Tree Classifier as in Figure 6, it is evident that the SVM model has higher precision. This indicates that the SVM model has a higher true positive rate and a lower false positive rate than the Decision Tree Classifier. Additionally, the SVM model also has a higher recall, indicating that it has a higher true positive rate and a lower false negative rate than the Decision Tree Classifier. Furthermore, the F-score, which takes both precision and recall, is also higher for the SVM model. This implies that the SVM model has higher values for both precision and recall, making it a better choice for this particular problem.

Figure 7 displays the ROC curve for both the SVM model and the Decision Tree Classifier. The AUC for the SVM model is closer to the top-left corner, indicating that the SVM model has a higher true positive rate and a lower false positive rate than the Decision Tree Classifier. So, we conclude that the SVM model is a better classifier than Decision Tree.

	Decision Tree Classifier	Support Vector Machine (SVM)
F-score	0.934	0.974
Recall	0.947	0.987
Precision	0.922	0.961

Figure 6: Evaluation matrices

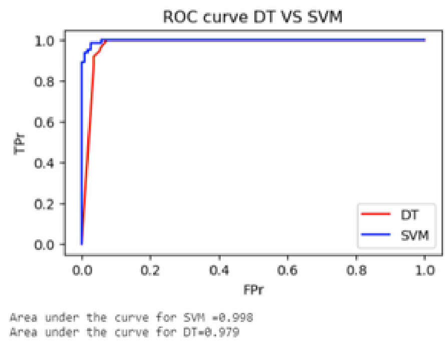

Figure 7: ROC & AUC for classifiers

| Conclusion

In summary, when choosing a machine learning model, it is essential to consider the specific problem and the characteristics of the data. Additionally, it is crucial to evaluate the performance of the model to ensure that it can make accurate predictions on new and unseen data. In the case of our problem, the SVM model appears to be the more suitable choice. This model has higher precision, recall, F-score, and AUC compared to the Decision Tree Classifier.

| References

- Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4). New York: Springer.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
- Mosaraf, M. P. (2023, January 21). Postpartum depression. Kaggle. https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression
- Sklearn.model_selection.RANDOMIZEDSEARCHCV. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- Bhandari, A. (2020, June 15). Guide to AUC ROC Curve in Machine Learning : What Is Specificity? Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/