Design a suitable pattern recognition model for classifying different water safety

Department of Computer Science and Artificial Intelligence, Jeddah University

CCAI-312: Pattern Recognition

By

Layali Alsolami

Bushra Dajam

Jana Osta

Dr. Safa Alsafari

june 3, 2023

Introduction

Water is a vital and essential resource for all living organisms on Earth. It covers about 71% of the Earth's surface and is found in oceans, rivers, lakes, and underground aquifers. Water plays a crucial role in various natural processes. However, having safe water remains a challenge for people around the world, especially the poor’s countries.

Without water, life on Earth would not be possible. That’s why we need it safe.

Problem Description

Unfortunately, there are communities throughout Africa that suffer from the lack of safe, clean water for drinking, cooking, and hygiene.

Data Description

First, the data set is a collection of data that is used to train the model. The solution of our problem is to find safe, healthy water. We chose the data set that describes the essential elements of different samples of different waters. Depending on these inputs which are the features? The output would be resulting in one of the binary classes which are *(Safe water class =1 , Not save water class = 0).*

The data contain **7996** ‘Rows’ as known as samples and **21** ‘columns’ including the features and the class label.

Methods

Initially, two hypotheses were developed to be verified in the balanced dataset using two algorithms Decision Tree and SVM (Support Vector Machine)

- If we have two hypotheses, one is that the water is safe and the other is water is not safe.
- Decision Tree is more effective than SVM in our dataset problem.

So, We use multiple evaluation matrix to evaluate and compare between the results such as accuracy , precision , recall , F1-measure, misclassification rate and AUC (Area under curve) .

Our goal is to try to improve our models performance by using the following approaches :

1- Different hyper parameter : we used GridSearchCV to tune params to achieve best performance for model.

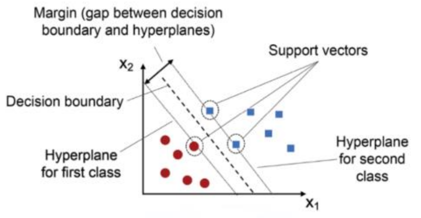
2- Preprocess the data: according to numerous searches. The best solution was to handle unbalanced dataset using a random oversample technique .

Decision Tree model:

we use Decision Tree for supervised learning for classification. It is like tree have root node. Internal nodes represent the features and branches represent the decision rules and leaf node must have same class (pure). splitting by using features . can use entropy to choose which feature should selected.

SVM model:

Support Vector Machine (SVM) is a type of supervised machine learning algorithm that is used for classification and regression analysis. It is a powerful and versatile algorithm that can be used for both linear and non-linear data classification. SVM works by finding the best possible boundary or hyperplane that separates the data points into different classes. The hyperplane is chosen in such a way that it maximizes the margin between the two classes, which helps to improve the accuracy of the model. SVMs are widely used in various applications such as image recognition, text classification, bioinformatics, and many more.


Margin (gap between decision boundary and hyperplanes)
Support vectors
x_2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x_1

Experiment and results

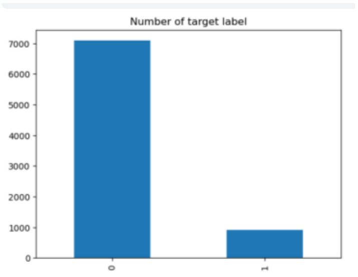
First, we started the preprocessing by apply the following steps :

- check for missing value
- remove all records that contain NUM!
- normalize dataset to make sure have same feature scaling
- use random oversample technique to handle unbalanced data

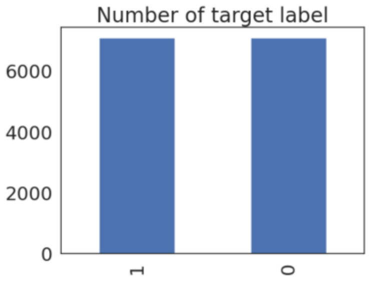
Random oversample:

This approach is used to increase the number of data points in the minority class by picking and duplicating random points from the minority class. The variance of the dataset is minimized because the points are chosen at random.

Before



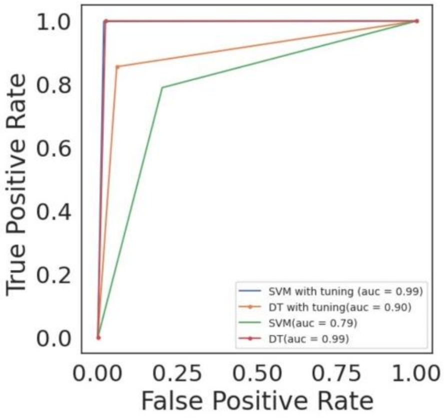
after



Then , we split data into train and test data with 80:20 ratio and then train the models. we got the following results from all the evaluation matrices.

	DT		DT_best		SVM		SVM_best	
Accuracy	0.986		0.898		0.7893		0.989	
precision	1.00	0.98	0.86	0.94	0.78	0.80	1.00	0.98
recall	0.98	1.00	0.94	0.86	0.80	0.78	0.98	1.00
F1-measure	0.99	0.99	0.90	0.90	0.97	0.97	0.99	0.99
misclassification	.0130		0.101		0.210		0.010	

As we can see in the table above. SVM with best hyperparameters has higher accuracy and less misclassification. Which, it mean that it is better than DT with best hyperparameters.
Here is ROC graph with AUC for each model, to show difference for models :



SVM line is top-left most which means has the best performance.

Discussion

As previously explained, we started with unbalanced data and did preprocessing process.
After building the model and separating the data for testing and training, we used Design Tree but the overfitting problem was appearing. Eventually, we decide to find the perfect hyperparameter, which is: {'criterion': 'gini', 'max_depth': 4, 'max_features': None, 'splitter': 'best'}. Then, we notice that the model is improved and give the best accuracy.

Basically, we set up the SVM model with the default hyperparameter which has the C value of 1, the gamma value of scale, and the kernel value of rbf. But as we expected the results was so inefficient. Then, we did the improvement by choosing the best hyperparameter for the SVM, which is Best hyperparameters is: {'C': 100, 'gamma': 1, 'kernel': 'rbf'}.
In the above figure, the two lines are close to each other, but SVM with tuning is the best because it is at the top and the left and has the largest AUC.

Conclusion

We based our project on data have the materials of the water. We explored the data, which in turn was not balanced. We decide to use oversampling technique to balance the data.
There were two hypotheses, one is that the water safe and the other water is not safe.
Two, Decision Tree is more effective than SVM in our dataset problem.

We conclude, that the second hypothesis was rejected. SVM was much efficient than Decision Tree according to the AUC numbers. Also, we successfully classified the water as if it was safe or not by our Decision Tree model. At the end, we hope that this project would help many African countries to dink a healthy safety water.

References

R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.

K. P. Bennett and J. A. Blue, "A support vector machine approach to decision trees," *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, Anchorage, AK, USA, 1998, pp. 2396-2401 vol.3, doi: 10.1109/IJCNN.1998.687237.

M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López, *Experiments with SVM to classify opinions in different domains*, Expert Systems with Applications (https://www.sciencedirect.com/science/article/pii/S0957417411008542)

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. D. L. F. (2018). *An empirical study on hyperparameter tuning of decision trees*. arXiv preprint arXiv:1812.02207.