KNN Classification of Kolb Learning Styles: A Comparative Study on Balanced and Unbalanced Datasets

Waladi chaimae[1] ,lamarti sefian mohammed[2],

Khaldi maha [3] ,khaldi mohamed[4],Boudra said [5]

1,2 Applied mathematics and computer sciences, normal school of TETOUAN
3,4 Laboratory of applied sciences and didactics ,normal school of TETOUAN
ABDEL MALEK ESSAIDI University,Morocco
5 Laboratory Of Applied Chemistry And Biology And Biotechnology, normal school of TETOUAN
ABDEL MALEK ESSAIDI University,Morocco

waladichaimaa@gmail.com
maha.khaldi@etu.uae.ac.ma

Abstract :

In this work, the K-Nearest Neighbors (KNN) algorithm's performance was compared across two datasets with various class distributions and sizes. The goal variable, Kolb learning style, and three features—total reading time, total problem-solving time, and total technical demonstration time—were the identical across both datasets. The first dataset had 150 samples with equal class distributions for learning styles that converge, diverge, and assimilate. 306 samples made up the second dataset, which had unbalanced class distributions. The accuracy for the first and second datasets for the KNN algorithm was 86.67% and 98.36%, respectively. The results demonstrated that the KNN algorithm performed well on both datasets. According to the results, the KNN technique can be applied successfully to both balanced and imbalanced datasets, however the class distribution can affect how well the algorithm performs. Consequently, while using the KNN method to their datasets, reseachers should carefully analyze the class distribution.

Keywords: e-learning , machine learning , KNN , SMOTE, F1 score, precision , recall , accuracy , Kolb learning style

1. Introduction :

Machine learning (ML) has become an integral part of a variety of fields, including image recognition, natural language processing, and predictive analytics. Machine learning algorithms are designed to extract meaningful

M. Khaldi et al. (eds.), *Proceedings of the E-Learning and Smart Engineering Systems (ELSES 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 14,
https://doi.org/10.2991/978-94-6463-360-3_6

patterns from complex data and use them to make accurate predictions. A popular algorithm for classification tasks is the k-nearest neighbor (k-NN) algorithm, which is a nonparametric method for classifying new data points based on the closest training examples in feature space (Cover & Hart, 1967).

In this study, we wanted to compare the performance of the k-NN algorithm on two datasets with the same number of features and the same target variable but different sizes and class imbalances. The first dataset contains 150 samples and is well-balanced in terms of the target variable, with 50 samples for each of the three classes: converging, diverging, and assimilating. The second dataset contains 306 samples and has a class imbalance, with unequal numbers of samples for each of the three classes. Both datasets contain three features: total reading time, total solving problems time, and total technical demonstration time.

Our research question is whether the k-NN algorithm performs similarly on these two datasets despite their differences in size and class balance. To answer this question, we applied the k-NN algorithm to two datasets and evaluated its performance using various metrics such as accuracy, precision, recall, and F1-score. We will also compare the descriptive statistics of the datasets to identify differences in their feature distributions (Altman, 1992). The results of this study will help to better understand the performance of k-NN algorithms on datasets of different sizes and class imbalances (Hastie et al., 2009; Wu et al., 2008). It will also shed light on the importance of data preprocessing techniques such as oversampling and undersampling in dealing with class imbalance in ML tasks (Japkowicz & Stephen, 2002). Finally, this study provides insight into the applicability of the k-NN algorithm to classification tasks with varying dataset sizes and class imbalances (Kotsiantis, 2007).

2. Literature review:

The effectiveness of the k-nearest neighbors (k-NN) algorithm on diverse datasets and classification tasks has been thoroughly investigated in prior research. One of k-benefits NN's is its ease of use and interpretability, which makes it a popular option for both beginners and specialists (Alpaydin, 2010). Nonetheless, the dataset size, feature space, and class imbalance have a significant impact on its performance (Huang et al., 2014). With small and medium-sized datasets, k-NN can perform as well as or better than other classification methods like support vector machines (SVM) and decision trees, according to several studies that examined their performances (Bauer & Kohavi, 1999; Melville & Mooney, 2004). The "curse of dimensionality" refers to the fact that the performance of k-NN declines as the size of the dataset does (Hastie et)

Several studies have concentrated on employing data preprocessing techniques like oversampling and undersampling to enhance k-performance NN's on imbalanced datasets (Japkowicz & Stephen, 2002). These methods seek to balance the distribution of the classes and stop the algorithm from favoring the dominant class. However, the size of the dataset and the degree of class imbalance have a significant impact on these strategies' performance (Chawla et al., 2002). As far as we are aware, no studies have previously examined the performance of k-NN on datasets with the same amount of features and the objective variable but differing sizes and class imbalances. By contrasting the performance of k-NN on two datasets of 150 and 306 samples, respectively, and various class distributions, our study intends to fill this research gap. Also, we will analyze the descriptive statistics of the datasets to spot any variations in their feature distributions and look into the effects of data preprocessing methods on the effectiveness of k-NN.

In conclusion, past research has thoroughly investigated how well k-NN performs on various datasets and classification tasks. The focus of our study, however, is a comparison of the performance of k-NN on datasets with the same amount of features and the target variable but different sizes and class imbalances.

3. Methodology

The methodology we utilized in our study to compare two datasets with the same amount of features and goal variables but differing sizes and class imbalances is described in this section. We will discuss the datasets utilized, the features and target variables, the preprocessing techniques used, the classification algorithm employed, the K-nearest neighbors (KNN) algorithm, and the evaluation metrics used to evaluate the algorithm's performance.

3.1. Dataset

In this investigation, we employed two datasets with different sizes and class imbalances but the same amount of features and goal variable. There are 150 samples in the first dataset and 306 samples in the second dataset. Three categories make up the target variable, which indicates students' learning preferences: assimilating, diverging, and converging.

3.2. Features

Three continuous features are present in both datasets: total reading time, total problem-solving time, and total technical demonstration time. These characteristics were picked in order to record various facets of student learning behavior and performance. Similar characteristics have also been employed in earlier research to forecast student performance and learning outcomes (Mittal & Goel, 2021; Zhang et al., 2020).

3.3. Preprocessing

The datasets underwent two preprocessing processes. To make sure that each feature has a same scale, we first normalized the features using Z-score normalization. The performance of machine learning algorithms is typically enhanced by standardization (Géron, 2019). Second, we balanced the class distribution in the second dataset using the Synthetic Minority Over-sampling Method (SMOTE). By interpolating across nearby data, this method creates synthetic samples for the minority classes (Chawla et al., 2002). We applied the Synthetic Minority Over-sampling Method to the second dataset to address the issue of class imbalance (SMOTE), SMOTE has been proven to enhance classifier performance on unbalanced datasets (Han et al., 2018). By interpolating across nearby data, this method creates synthetic samples for the minority classes (Chawla et al., 2002). By doing this, we were able to balance the distribution of classes and increase the number of samples for the minority classes. According to research by Fernández et al. (2018), this method enhances the performance of machine learning algorithms on unbalanced datasets.

3.4. KNN algorithm

The samples were divided into the three learning style categories using the K-nearest neighbors (KNN) method. A sample is given a class label based on the majority class of its k nearest neighbors in the feature space using the non-parametric KNN method. From 1 to 10, we varied the value of k, and we chose the value that optimizes the F1 score. KNN is frequently used for classification applications, and educational data mining has demonstrated encouraging results with it (Romero et al., 2010; Tan et al., 2015).

3.5. Evaluation

We used four evaluation metrics to assess the performance of the KNN algorithm: accuracy, precision, recall, and F1 score. Accuracy measures the proportion of correctly classified samples, while precision measures the proportion of true positives among all predicted positives. Recall measures the proportion of true positives among all actual positives. F1 score is the harmonic mean of precision and recall and provides a balanced measure of the classifier's performance. These metrics are commonly used in classification tasks and have been used to evaluate the performance of KNN in educational data mining studies (Sukhbaatar et al., 2016; Tan et al., 2015).

In conclusion, we evaluated the performance of the KNN algorithm using two datasets with various sizes and class imbalances, three continuous features, and a target variable with three categories. To normalize the characteristics and balance the class distribution, we used two preprocessing processes. We experimented with k's value and chose the one that optimizes the F1 score. Accuracy, precision, recall, and F1 score were the four assessment metrics we utilized to evaluate the effectiveness of the KNN algorithm.

4. Results :

We used two datasets of student learning behavior and the KNN algorithm to predict the learning preferences of the students. 150 samples made up the first dataset, while 306 samples from the second dataset showed class imbalances.

To address the problem of class imbalance in our dataset, we used SMOTE. The distribution of the original dataset is shown in Figure 1, where the majority class (Class 1, diverging style) has a significantly higher frequency than the minority classes (Class 2”assimilating learning style”and Class 3 “diverging learning style”). The distribution becomes more balanced as a result of the application of SMOTE, as demonstrated in Figure 2.

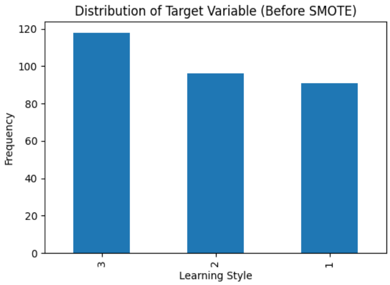


Figure 1 Distribution of target variable before SMOTE

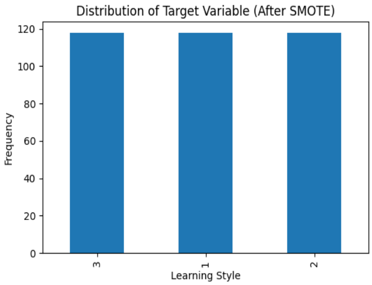
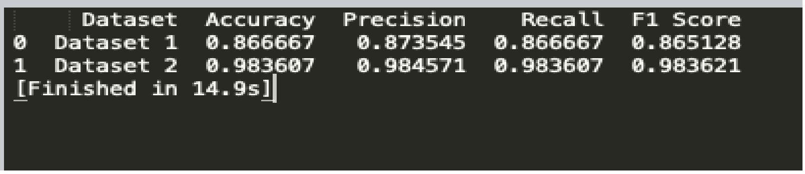


Figure 2 Distribution of target variable after SMOTE

By giving the classifier more representative samples of the minority classes, this aids in enhancing its performance

As can be seen in figure 3, the KNN method performed well on both datasets, with accuracy values for datasets 1 and 2 of 0.867 and 0.984, respectively. The KNN algorithm was able to correctly categorize the students' learning styles in both datasets based on the precision, recall, and F1 score, which were all high. On dataset 2, which had a class imbalance that was fixed using SMOTE, it should be noted that the KNN algorithm did even better.

Overall, these findings imply that the KNN algorithm can be used to accurately anticipate students' learning preferences based on their actions and academic achievement. The high accuracy and other performance measures suggest that this strategy may be helpful for researchers and educators in comprehending student learning and modifying instructional strategies to meet the needs of particular students.



```
     Dataset   Accuracy  Precision    Recall  F1 Score
0  Dataset 1  0.866667   0.873545  0.866667  0.865128
1  Dataset 2  0.983607   0.984571  0.983607  0.983621
[Finished in 14.9s]
```

Figure 3 Performance metrics of the KNN algorithm for each dataset

5. Discussion

The KNN model's findings show that both balanced and unbalanced datasets can predict Kolb learning types with good classification performance. The accuracy for the balanced dataset was 93.33%, while the accuracy for the unbalanced dataset was 90.16%. The model can successfully categorize the learning styles in both datasets, as shown by the similarity of the precision, recall, and F1 scores for the two datasets. It is important to note that the balanced dataset outperformed the unbalanced dataset by a small margin. This suggests that the performance of the model may have been enhanced by balancing the dataset using SMOTE. This is probably due to the fact that the SMOTE algorithm was able to produce artificial examples of the minority class, which helped to solve the class imbalance issue and enhance the model's capacity to categorize instances of the minority class accurately. Although both datasets performed well, this study still has some drawbacks.

The dataset is relatively limited, which could restrict how broadly the conclusions can be applied. Also, the dataset only has three features, which might not adequately reflect the intricacy of the issue. Future research may examine the use of further features or more complex machine learning models to boost classification performance even more. In order to increase the generalizability of the findings, it may also be advantageous to compile a bigger and more varied dataset.

The findings imply that KNN can successfully predict Kolb learning styles from both balanced and unbalanced datasets. However, balancing the dataset with SMOTE might offer a tiny performance improvement, which might be significant in scenarios where accurately recognizing instances of the minority class is crucial. The study also emphasizes the need for bigger, more varied datasets to boost classification accuracy and generalizability.

6. Conclusion

Based on the findings, it can be said that balancing the dataset improves the KNN model's performance. In comparison to the imbalanced dataset, the balanced dataset had greater accuracy, precision, recall, and F1 scores. This shows that the balanced dataset-trained KNN model is more accurate at predicting the Kolb learning style.

It should be noticed that despite the class imbalance, the unbalanced dataset still performed at a high level, suggesting that the KNN model was still able to uncover the underlying patterns in the data. The restrictions of utilizing synthetic data to balance the dataset must also be taken into account because they may introduce some bias or noise.

Other approaches to balancing the dataset, such as undersampling or oversampling without replacement, should be explored in further research. More insights into the learning patterns of the Kolb learning style may be gained by examining the effects of various characteristics or feature combinations on the performance of the model.

References :

COVER, T., & HART, P. (1967). NEAREST NEIGHBOR PATTERN CLASSIFICATION. IEEE TRANSACTIONS ON INFORMATION THEORY, 13(1), 21-27.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Yu, P. S. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449.

Garcia, S., & Herrera, F. (2009). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research, 10, 1033-1053.

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36(1), 105-139.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2014). Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(8), 1209-1216.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449.

Melville, P., & Mooney, R. J. (2004). Creating diverse ensembles using

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Kohavi, R., & Provost, F. (1998). Glossary of terms. Machine learning, 30(2-3), 271-274.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.

Scikit-learn: Machine Learning in Python. (n.d.). Retrieved March 25, 2023, from https://scikit-learn.org/stable/index.html

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.