Mathematical Modeling of Monetary Poverty by K-Nearest Neighbors Algorithm

EL AACHAB Yassine[1] and , KAICER Mohammed[2]

[1] Laboratory of Analysis Geometry and Applications Department of Mathematics, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco
[2] Laboratory of Analysis Geometry and Applications Department of Mathematics, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco
y.elaachab@gmail.com

Abstract. The K-Nearest Neighbors (KNN) algorithm is used in this study to categorize households into poor and non-poor groups using information from the National Household Consumption and Expenditure Survey 2013/2014 which is organized every ten years by the high commissioner for planning in morocco. The KNN algorithm is a nonparametric instance-based method that bases predictions on how similar the target instance and its neighboring examples are. It is possible to effectively categorize families and identify those who are living in poverty by training the KNN model on a dataset comprising pertinent demographic and socioeconomic traits. This study intends to investigate the KNN algorithm's efficiency in classifying households and its ability to offer insightful information on efforts to combat poverty.

Keywords: Machine learning, Classification, Prediction, K-Nearest Neighbors, Household Poverty.

1 Introduction

Due to their capacity to evaluate enormous datasets and generate precise predictions, machine learning algorithms have attracted a lot of interest lately[1]. These algorithms have been used in a variety of fields, including social sciences, finance, healthcare, and marketing. Machine learning algorithms offer a viable method for dividing families into poor and non-poor groups based on pertinent socioeconomic characteristics in the context of poverty studies[2].

The K-Nearest Neighbors (KNN) technique is one of many such algorithms that are frequently employed for classification problems[3]. A non-parametric technique called KNN is based on the idea that comparable objects frequently are members of the same class. Based on the class labels of its k nearest neighbors in the feature space, it categorizes an unlabeled instance. KNN has been used to good effect in a number of fields, including anomaly detection, picture recognition, and recommendation systems[4].

For organizations and policymakers aiming to reduce poverty, the categorization of homes into poverty and non-poor groups is crucial[5]. It is possible to tailor interventions, allocate resources, and make policy decisions to address the unique needs and challenges encountered by these households when poverty-stricken households are accurately identified. In particular, KNN offers the potential to increase the precision and effectiveness of poverty classification, resulting in more potent measures for reducing poverty[6].

2 Mathematical formulation of KNN

The neighborhood $C_k(\hat{y},\{y\})$of a new sample $\hat{y}$ is evaluated based on distance which can be treated as a hyper-parameter of the algorithm. Common choices are the minkowski distance formed as :

$$d(\hat{y}),y) = (\sum_{l=1}^{k}|\hat{y} - y_i|^n)^{\frac{1}{n}} \quad 1$$

Where n is the degree of choice for the minkowski distance.For the case n=2 Euclicean distance is defined while other choices might be desirable for certain settings (Manhattan, Hamming).

The chosen distance is used to retrieve the K nearest neighbors of the sample majority class among the dataset samples.in the brute force approach no learning occurs.The method relies uniquely on the online evaluation of the samples x_l , based on the distance function of choice. Thus given the neighborhood $C_k(\hat{y},\{y\})$, the online objective can be formulated as :

$$\hat{z} = \max_{z\in Z} \sum_{z_i\in(y_i,z_i)}^{C_k(\hat{y},\{y\})} \zeta(z_i - z) \quad 2$$

With $\zeta(z_i - z)$ being the Dirane data function acting as a counter :

$$\zeta(z_i - z) = \begin{cases} 1 & if\ z_i = z \\ 0 & else \end{cases} \quad 3$$

3 Data and software

The database used in this study to measure the performance of the KNN method for predicting and classifying household monetary poverty is a database from the High Commission for Planning, from its survey called the National Survey on household consumption and expenditure, 2014 organized every ten years. To manipulate the data, we opted for the R studio software.

4 Results and discussion

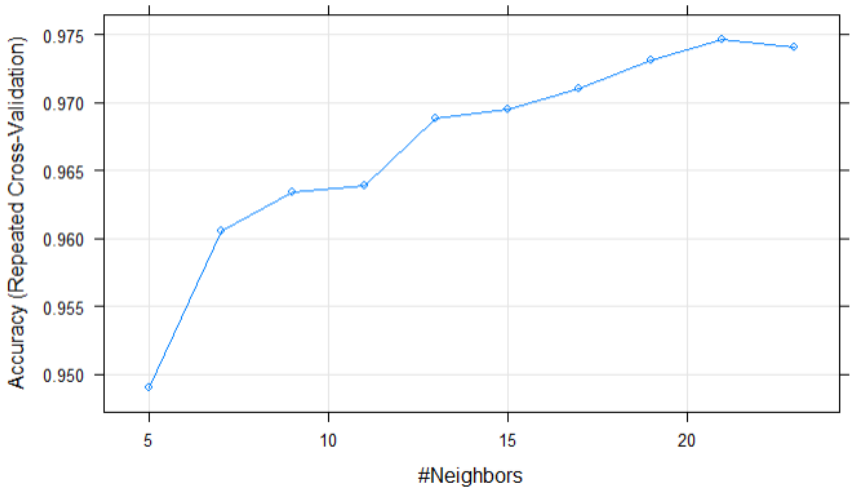

Fig. 1. The variation of the value of the precision according to the variation of the value of K.

Based on the results, it seems that the KNN algorithm achieves high accuracy for different values of k. The accuracy increases as the value of k increases, up to a certain point. Additionally, the Kappa coefficient, which measures the agreement between predicted and actual class labels beyond chance, also shows a generally increasing trend with increasing k.

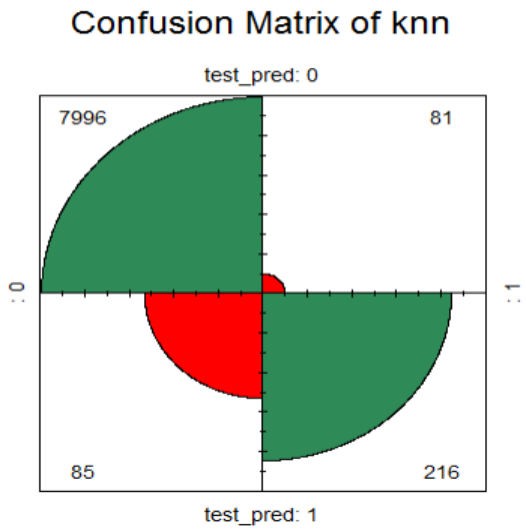

Fig. 2. Confusion Matrix of KNN.

The model achieved a high number of true negatives (7,996), indicating a strong ability to correctly identify non-poor households. This suggests that the model accurately identified households that are not in poverty, which is essential for avoiding misallocation of resources to non-poor households.

However, there were a relatively small number of false positives (81), indicating instances where the model wrongly predicted non-poor households as poor. This suggests that some non-poor households were misclassified as being in poverty, which could potentially result in misdirected assistance to those households.

The false negatives (85) represent instances where the model incorrectly classified poor households as non-poor. This misclassification is particularly important, as it may lead to the exclusion of genuinely impoverished households from receiving the assistance they require.

On the positive side, the model achieved a decent number of true positives (216), indicating its ability to correctly identify households living in poverty.

Overall, while the model demonstrated a reasonable ability to predict household poverty, there is room for improvement, particularly in reducing false positives and false

negatives. Further analysis, model refinement, or incorporating additional relevant features may help enhance the model's performance in accurately classifying households into poor and non-poor categories.

Table 1. Metrics of KNN .

Metric	Value
Accuracy	0,98
Precision	0,73
F1 Score	0,99
Kappa	0,71
McNemar's Test	0,82
Sensitivity	0,99
Specificity	0,73
Pos Pred Value	0,99
Neg Pred Value	0,72
Prevalence	0,96
Detection Rate	0,95
Detection Prevalence	0,96
Balanced Accuracy	0,86

The classification of households into poor and non-poor categories was conducted using a machine learning model, and the evaluation metrics provide insights into its performance. The model achieved an accuracy of 98.02%, indicating that it correctly classified the majority of instances in the dataset. Precision, which measures the proportion of correctly predicted positive instances out of all instances predicted as positive, was found to be 73%. This suggests that out of all instances predicted as poor households, 73% were indeed correctly classified. The F1 score, a balanced measure of precision and recall, was calculated to be 0.9897, indicating the model's ability to achieve a high overall performance. Furthermore, the Kappa coefficient, which measures the agreement between predicted and actual class labels beyond chance, was found to be 0.7121. The sensitivity (or recall) value of 98.95% indicates that the model correctly identified a large proportion of actual positive instances (poor households). On the other hand, the specificity value of 72.73% suggests that the model accurately classified a moderate proportion of actual negative instances (non-poor households). Additional evaluation metrics, such as positive predictive value, negative predictive value, prevalence, and balanced accuracy, provided further insights into the model's precision, prevalence of the positive class, and overall accuracy. Overall, the evaluation metrics demonstrate the model's ability to accurately classify households into poor and

non-poor categories, with a high level of precision and a strong balance between identifying positive and negative instances.

Conclusion

In conclusion, the KNN classification model applied to classify households into poor and non-poor categories demonstrated strong performance and yielded promising results. The model achieved an accuracy of 98.02%, indicating its ability to correctly classify the majority of instances. The precision of 73% suggests that when predicting positive instances (poor households), the model had a relatively high rate of correct classifications. The F1 score of 0.9897, which combines precision and recall, further emphasizes the model's overall effectiveness. The Kappa coefficient of 0.7121 indicates a substantial agreement between predicted and actual class labels beyond chance.

Furthermore, the sensitivity (recall) of 98.95% implies the model's ability to correctly identify a large proportion of actual positive instances (poor households). The specificity of 72.73% demonstrates the model's ability to accurately classify non-poor households. These evaluation metrics collectively reflect the model's balanced performance in identifying both positive and negative instances.

Based on these results, it can be concluded that the KNN classification model shows promise for accurately categorizing households into poor and non-poor categories. However, it is important to note that further analysis, experimentation, and refinement may be necessary to address potential limitations and enhance the model's performance. Future research could focus on exploring additional features, applying feature selection techniques, or considering alternative algorithms to improve the classification accuracy and robustness of the model.

References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
2. Alim, M. A., Kim, H. G., & Lee, Y. K. (2017). A comparative analysis of K-Nearest Neighbor and Support Vector Machine for land cover classification using satellite images. Sustainability, 9(10), 1840.
3. Ghani, R. A., & Muthu, R. (2015). Anomaly detection using k-nearest neighbor classification algorithm. Procedia Computer Science, 47, 51-57.
4. N'Guessan, A., & Bro, P. (2019). Poverty classification from remote sensing data using machine learning. Remote Sensing, 11(1), 91.
5. Ravindran, A., Zafarani, R., & Liu, H. (2012). The K-Nearest Neighbors Algorithm. In Social Media Mining (pp. 147-171). Cambridge University Press.
6. World Bank. (2021). Poverty Overview. Retrieved from https://www.worldbank.org/en/topic/poverty/overview