ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY517

Investigating the Impact of Sample Size on the Performance of the k-NN Algorithm

Zara Wong
University of Colorado Boulder,
3100 Marine St. Boulder, CO
80309-0065 USA

Abstract:- The k-Nearest Neighbour (k-NN) algorithm is a simple and intuitive classification algorithm used for pattern recognition and classification tasks. This research paper aims to address a gap in literature by exploring the relationship between sample size and the performance of the k-Nearest Neighbour (k-NN) algorithm. Through intensive experimental analysis of secondary data, we investigate how varying sample sizes influence the algorithm's classification accuracy, computational efficiency, and generalization capabilities. Our findings reveal that an ideal scope for sample sizes is >190, with minimal differing results beyond that point. The maximum of the graph is 340, suggesting it to be the optimal value for ideal accuracy for this training model and scope. These results contribute to a deeper understanding of the proper application of the k-NN. These findings contribute to a deeper understanding of the complex interplay between sample sizes and k NN algorithm performance, aiding practitioners in making informed decisions when employing this method in real-world applications, and suggest the ideal value for sample size.

Keywords:- *Systems Software; Algorithms; Machine Learning; k-Nearest Neighbour; Sample Size.*

I. INTRODUCTION

The k-Nearest Neighbour (k-NN) algorithm stands as a foundational method in pattern recognition and classification. Its simplicity and potential for high accuracy have contributed to its widespread use across diverse domains. However, the impact of sample size on its performance is an area that could still remain to benefit from deeper exploration.

This paper aims to address this gap in literature by delving into the complex relationship between sample size and k-NN algorithm performance. The hypothesis is that the optimal sample size exists within a range that balances the model's ability to capture underlying patterns while avoiding overfitting or underfitting. Understanding the influence of sample size on k-NN algorithm performance holds practical significance. In the era of large datasets, selecting an appropriate sample size becomes crucial. The objective of the experimental analysis is to investigate how varying sample sizes impact the algorithm's classification accuracy, computational efficiency, and generalization capabilities.

II. LITERATURE REVIEW

The k-NN algorithm, introduced in 1951, remains a relevant classification technique due to its simplicity and effectiveness. However, few studies have thoroughly examined the role of sample size in its performance. This paper will focus on the IRIS dataset. Through a comparison of results between varying sample sizes, the effect of sample size on k-NN classification performance will be determined[1].

Experts in the field have highlighted the risks of small sample sizes, which can lead to overfitting as the algorithm learns noise rather than true patterns. Conversely, large sample sizes can cause oversimplification, leading to underfitting. The optimal sample size is thus a critical parameter that requires thorough investigation[2].

Related works noted diminishing returns in performance with smaller sample sizes on specific dataset[2],Additionally, they suggest that the effectiveness of dimensionality reduction techniques might be influenced by variations in sample size.

These findings highlight the need for a comprehensive analysis of the relationship between sample size and various dimensions of k-NN algorithm performance[3].

This research extends this body of knowledge by systematically exploring how sample size impacts not only classification accuracy but also computational efficiency, generalization capabilities, and potential biases. By providing a holistic understanding, we aim to guide practitioners in selecting appropriate and optimal sample sizes for their specific applications.

➢ *Machine Learning*

Machine learning is a subset of artificial intelligence that empowers computers to learn from data and improve their performance on specific tasks without being explicitly programmed. Instead of following fixed rules, machine learning algorithms use patterns and information present in training data to make predictions or decisions[4]. By recognizing complex relationships within data, machine learning models can generalize their knowledge to make accurate predictions on new, unseen data. This process involves iterative refinement as the model learns from mistakes and adapts to changing data patterns. Machine learning finds applications in diverse fields such as image recognition, natural language processing, recommendation systems, and more, driving advancements across various industries[5]. k-NN is one of the most simple yet essential machine learning algorithms, falling under the category of supervised learning.

➢ *The IRIS Dataset*

One of the datasets this investigation was based upon was the IRIS dataset. This is a cleaned dataset introduced by statistician Ronald Fisher that serves as a test case for multiple statistical classification techniques. It consists of 50 different samples from 3 species, with a total of 150 records under 5 attributes[6]. A few examples from the dataset are shown in Table 1.

Table 1 Examples from the IRIS Dataset

#sepal_length	#sepal_width	#petal_length	#petal_width	species
5.1	3.5	1.4	0.2	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
6.3	3.3	6	2.5	Iris-virginica

➢ *Synthetic Dataset*

The second dataset this investigation was based upon in order to provide a greater range and easier manipulation than the IRIS dataset was a systematically generated Synthetic Dataset with values as shown in Table 2.

Table 2 Parameters for the Synthetic Dataset

	Class 1	Class 2
μ	4	-4
σ	2	10
d^1	2	2

III. METHODS

➢ *k-Nearest Neighbour*

The k-Nearest Neighbour (k-NN) algorithm is a simple and intuitive classification algorithm used for pattern recognition and classification tasks[7]. At its core, the algorithm assigns a class label to a new data point based on the class labels of its k nearest neighbors in the training dataset. The class label that appears most frequently among these k neighbors is assigned to the new data point. The decision-making process can be represented using the following equation:

$$C(x) = arg\ max_{c_j}\ \Sigma_{i=1}^{k} I(y_i = c_j\ \ )$$

$C(x)$: The predicted class label for the unlabeled instance *x*.

c_j: Represents a specific class label in the dataset. For example, in a binary classification problem, there would be two classes, class 0 or class 1, with $c_j$representing one of them.

y_i: The class label of the *i*-th nearest neighbor of the unlabeled instance *x*.

k: Input parameter signifying the number of nearest neighbors to consider when making a prediction

I(y_i=c_j):An indicator function that returns 1 if y_i=c_j, and 0 otherwise.

In order to determine the closest groups/nearest points for the KNN equation, distance metrics such as the Euclidean distance must be calculated. This can be visualized simply as the length of the line joining the point *x* in consideration to the surrounding points, in order to find the *k* nearest neighbors. Euclidean distance is calculated by the following equation:

$$d(x,y) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2}$$

The simple logic of the k-NN algorithm can be represented visually as demonstrated in Figure 1:

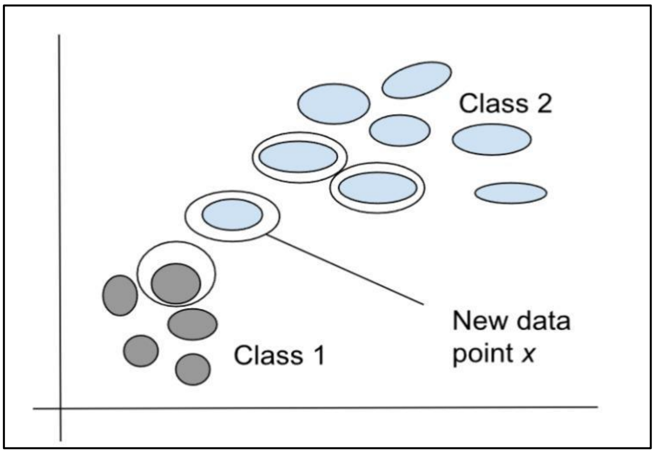

Fig 1 Visual Representation of k-NN

➢ *Methodology*

Both the IRIS dataset and synthetic data were used in order to accurately test the varying performance of the kNN algorithm with different sample sizes. The model went through a training period with a predetermined *k* value of 3, throughout which it was fed training data with corresponding labels. The model was subsequently prepared to recognize underlying patterns and perform calculations with unseen data, as exemplified in Figure 2.

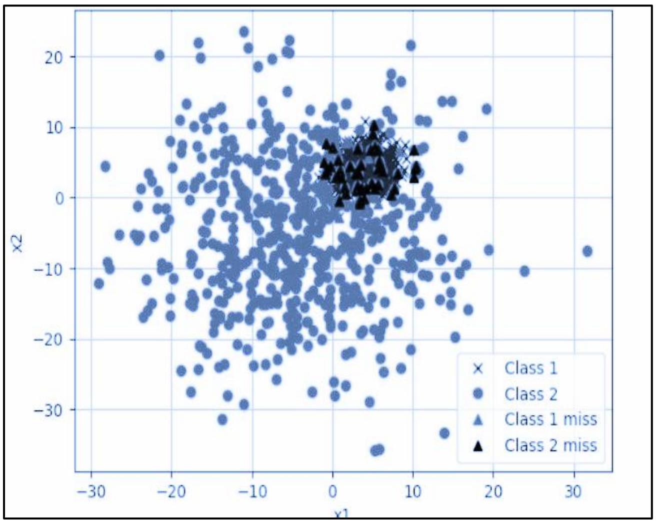

Fig 2 Example Test Set Classification Results

To investigate the impact of sample size on the k-NN algorithm, the sample size was systematically varied on both the synthesized data and the IRIS dataset, ranging from small subsets to the complete dataset. The data was split into training and testing sets. In order to ensure reliability the experiment was conducted a numerous amount of times. The accuracy of the model with different sample sizes was iteratively calculated and documented on a graph and systematically narrowed down in order to determine the optimal values for sample size.

IV. RESULTS AND DISCUSSION

The average was taken iteratively from 10 calculations with sample sizes from 5, 10, to 100 with steps of 10. However, though seemingly efficient with a run time of >0.04 the results proved to be too inconsistent. The average was then taken from 1000 sample sizes from 1000 differing calculations, where it was then determined that it would be most efficient to include larger sample sizes in order to systematically narrow them down to reach optimal size as shown in Figure 3.

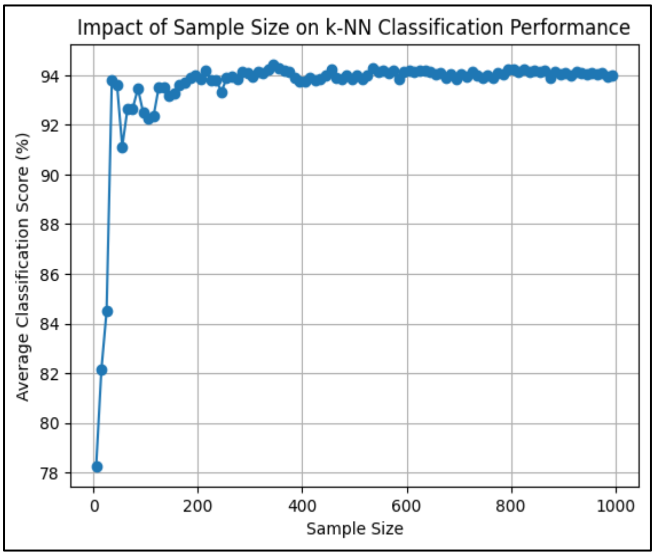

Fig 3 Run 1 with a runtime of >14.05

As exemplified, with a sample size of 5>120, the results are significantly more randomized across the plot with error rates spanning from ~22% to ~6%. However, the accuracy begins to rise at an increasingly logical rate at sample size 50, beginning to form a positive function with a linear trendline $7.82E-3*x+92.5$ from 120<190 with a moderate strength of $R^2 = 0.55$, as demonstrated in Figure 4:

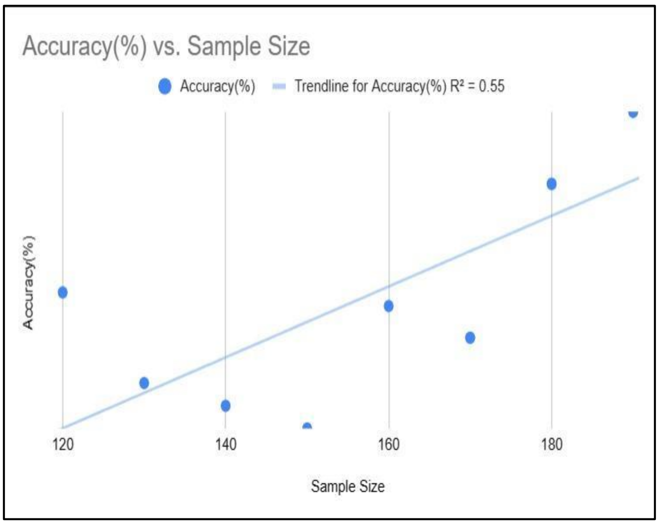

Fig 4 Linear Trendline of 120<190

Beyond this function, <190 sample sizes face little change, residing along the line of ~92% accuracy. Therefore, above 190, sample size plays little effect on the accuracy of the k-NN algorithm and is predominantly either beneficial or detrimental to the results. It is worth noting that at 340 the graph experiences a minor surge upwards signifying that 190 is the ideal sample size for scope >0, definitively 0<1001 for this training model.

V. CONCLUSION

These results suggest that an ideal scope for sample sizes is >190, with minimal differing results beyond that point. The maximum of the graph is 340, suggesting it to be the optimal value for ideal accuracy for this training model and scope. These findings help to contribute to a deeper understanding of the complex interplay between sample size and k-NN algorithm performance, aiding practitioners in making informed decisions when employing this method in real-world applications, and suggesting the ideal value for sample size. Future research in this area could focus on diverse algorithms, comparisons of the impact of sample size on k-NN to other algorithms, the addition of varied *k*-values, and investigating whether optimal sample sizes differ when tampering with other hyper-parameters such as distance metrics or data normalization techniques.

ACKNOWLEDGEMENTS

I would like to extend a thank you to the team at Indigo Research for providing me with the resources needed to execute this project.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY517

REFERENCES

[1]. Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets. *SN Applied Sciences*, *1*(12), 1559. https://doi.org/10.1007/s42452-019-1356-9

[2]. Trivedi, U. B., Bhatt, M., & Srivastava, P. (2021). Prevent Overfitting Problem in Machine Learning: A Case Focus on Linear Regression and Logistics Regression. In P. K. Singh, Z. Polkowski, S. Tanwar, S. K. Pandey, G. Matei, & D. Pirvu (Eds.), *Innovations in Information and Communication Technologies (IICT-2020)* (pp. 345–349). Springer International Publishing. https://doi.org/10.1007/978-3-030-66218-9_40

[3]. Maia Polo, F., & Vicente, R. (2023). Effective sample size, dimensionality, and generalization in covariate shift adaptation. *Neural Computing and Applications*, *35*(25), 18187–18199. https://doi.org/10.1007/s00521-021-06615-1

[4]. Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Current Reviews in Musculoskeletal Medicine*, *13*(1), 69–76. https://doi.org/10.1007/s12178-020-09600-8

[5]. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

[6]. Mithy, S., Hossain, S., Akter, S., . U. H., & Sogir, S. (2022). *Classification of Iris Flower Dataset using Different Algorithms*. *9*, 1–10. https://doi.org/10.26438/ijsrmss/v9i6.110

[7]. Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-10358-x