A Comparative Analysis of K-Nearest Neighbors and Naive Bayes Algorithms for Classifying Abnormal and Normal Spine Datasets

Y. Naga Vyshnavi

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract*—** *The accurate classification of spine datasets into "Abnormal" and "Normal" classes is crucial for early diagnosis and effective treatment planning in orthopedic medicine. In this paper, we present a comparative study of two popular machine learning algorithms, K-Nearest Neighbors (KNN) and Naive Bayes, applied to a spine dataset. The objective is to determine which algorithm performs better in terms of accuracy and suitability for this specific classification task. We evaluate both methods using a dataset consisting of 310 spine samples, labeled as either "Abnormal" or "Normal." Our results demonstrate that Naive Bayes outperforms KNN, achieving an accuracy of 89%, compared to KNN's accuracy of 87%. We also discuss the implications of these findings and highlight potential areas for further research in spine dataset classification.*

I. INTRODUCTION

AI is a subfield of computerized reasoning that spotlights on the improvement of calculations and models that empower PCs to learn and pursue expectations or choices without being expressly customized. It includes the investigation of measurable and computational strategies that permit machines to gain examples and concentrate experiences from information [1].

Exact expectation assumes a crucial part in different applications, like clinical finding, monetary estimating, and client conduct examination. Multi-facet Perceptron and Strategic Relapse are broadly utilized calculations in the field of prescient demonstrating [2][3]. This paper plans to direct a thorough report to look at their exhibition and dissect their materialness for various expectation errands. The review explores various assessment measurements to evaluate the forecast exactness and viability of the two calculations on vote dataset.

II. SUPERVISED LEARNING

Regulated learning is a particular sort of AI task where the calculation gains from named information. In regulated learning, the preparation dataset comprises of information (includes) and comparing yield marks. The objective is to gain proficiency with a planning capability that can foresee the result marks for new, concealed input information.

The preparation cycle in regulated learning includes giving the calculation a bunch of marked models and permitting it to get familiar with the basic examples or connections between the info highlights and result names. The calculation iteratively changes its interior boundaries to limit the contrast between its anticipated results and the genuine names in the preparation information. This cycle is regularly directed by a goal capability, for example, limiting the mean squared blunder or augmenting the probability of the noticed marks [3][4][6].

When the regulated learning calculation has been prepared on the marked information, it tends to be utilized to make forecasts on new, inconspicuous information by applying the got the hang of planning capability. The calculation takes the info elements of the new information as information and produces anticipated yield names as result [6][7].

Managed learning calculations can be additionally arranged into two principal types: relapse and order. Relapse calculations are utilized when the result variable is constant, and the objective is to foresee a mathematical worth [1][2]. For instance, anticipating the cost of a house in light of its elements. Arrangement calculations, then again, are utilized when the result variable is clear cut or discrete, and the objective is to allot new information to predefined classes or classifications. For instance, characterizing messages as spam or non-spam in light of their substance.

III. METHODOLOGY

AI is a logical procedure where the PCs figure out how to tackle an issue, without expressly program them. Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and models that enable computers to learn

and make predictions or decisions without being explicitly programmed. It involves the study of statistical and computational methods that allow machines to learn patterns and extract insights from data.

Accurate prediction plays a vital role in various applications, such as medical diagnosis, financial forecasting, and customer behavior analysis. This paper aims to conduct a comprehensive study to compare their performance and analyze their applicability for different prediction tasks. The study investigates multiple evaluation metrics to assess the prediction accuracy and effectiveness of both algorithms on Spine dataset.

3.1 K Nearest Neighboor (KNN)

KNN classifier is a directed languid classifier which has nearby heuristics. Being a lethargic classifier, involving this for expectation progressively is troublesome. The choice limits you accomplish with K-NN are significantly more perplexing than any choice trees, hence getting a decent grouping [3][4]. At the point when you are taking care of an issue which straightforwardly focusses on finding similitude between perceptions, K-NN improves due to its inborn nature to locally streamline. This is likewise a flipside on the grounds that, exceptions can essentially kill the presentation. Furthermore, K-NN is probably going to overfit, and thus changing 'k' to boost test set execution is the best approach.

K-Closest Neighboor is a strategy for grouping objects in light of learning information that is the nearest distance or has the most trademark similitudes with the item close or far neighbors are generally determined by Euclidean distance [4].

Steps-moves toward ascertain the K-NN calculation:

- Decide the worth of k.
- Ascertain the square of Euclid distance (question occasions) of each item against the preparation information.
- Sort objects-these items into bunches that have the littlest Euclid distance.
- Gathering class Y marks (Closest Neighborhood characterization)

3.2 Naïve Bayes

Naive Bayes is an energetic learning classifier and it is a lot quicker than K-NN. Subsequently, it very well may be utilized for expectation continuously. Commonly, email spam separating utilizes Credulous Bayes classifier [5][6]. It takes a probabilistic assessment course and creates probabilities for each class. It expects contingent freedom between the elements and utilizations a most extreme probability speculation. The most awesome aspect with this classifier is that, it learns over the long run. In a spam separating task, the sort of spam words in email develops after some time. Similarly, the classifier likewise works out likelihood gauges for the recently happening spam words in a "sack of words" model and ensures it performs well. This element of the classifier is because of the innate idea of the calculation being generative and not discriminative.

IV. EXPERIMENTAL RESULTS

The investigations have been coordinated by using Python programming tongue. The Python Scikit-learn is a pack for data portrayal, gathering and portrayal. The Spine dataset used in this review was procured from the UCI ML vault data set [7]. In this vote dataset there are 310 cases and 12 elements recorded and 2 class marks. The standard dataset is distributed two sets one for preparing (70%) and one more set for testing (30%).

4.1 Results:

We conducted a comprehensive experiment using the K-Nearest Neighbors (KNN) and Naive Bayes algorithms to classify a spine dataset into two classes: "Abnormal" and "Normal." The dataset comprises 310 samples, with 210 samples labeled as "Abnormal" and 100 samples labeled as "Normal." The results obtained from our experiments are as follows:

Table-1
Experimental Results

Algorithm	Accuracy	Precision	Recall
K-Nearest Neighbours	87.24	87.2	87.3
Naïve Bayes	89.46	89.5	89.5

4.2 Discussions:

The comparative analysis revealed that the Naive Bayes algorithm outperformed the K-Nearest Neighbors (KNN) algorithm in classifying the spine dataset into "Abnormal" and "Normal" categories. Naive Bayes achieved an accuracy of 89.46%, which is 2% higher than KNN's accuracy of 87.24%.

The higher accuracy of Naive Bayes can be attributed to its underlying probabilistic nature, which allows it to handle categorical data effectively. On the other hand, KNN is sensitive to the choice of distance metric and the number of neighbors, which might not be well-suited for the given dataset.

However, it is essential to consider the trade-offs between precision and recall for both algorithms. While Naive Bayes demonstrated higher accuracy, it might have different trade-offs depending on the application. For instance, in medical diagnoses, it is crucial to have high recall (sensitivity) to minimize false negatives, as missing a positive case (e.g., "Abnormal" spine) can have severe consequences.

Further investigation is warranted to explore the interpretability and robustness of the selected models. Additionally, employing other machine learning techniques, such as Decision Trees or Support Vector Machines, could provide valuable insights into their performance for spine dataset classification.

V. CONCLUSION

In conclusion, our study demonstrates that Naive Bayes is a promising approach for the classification of spine datasets, outperforming the K-Nearest Neighbors algorithm in terms of accuracy. However, the choice of the most appropriate algorithm should consider specific application requirements and performance metrics like precision and recall. The findings of this study could contribute to the development of more accurate and reliable models for spine dataset classification, ultimately enhancing medical decision-making and patient care.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G Ravi Kumar, K Venkata Sheshanna and G Anjan Babu, “Sentiment analysis for airline tweets utilizing machine learning techniques”, International Conference on Mobile Computing and Sustainable Informatics, PP:791-799, Publisher:Springer, Cham, 2020

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”,2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] UCI machine learning repository. http://archive.ics.uci. edu/ml/