Evaluating the Performance of Naive Bayes and K-Nearest Neighbor Models for Predictive Analysis

Ramavath Venkata Siva Kalyan Lal Naik

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Ongoing studies in supervised learning highlight the effectiveness of Naive Bayes classifiers. However, the question arises whether less restrictive classifiers perform better. This paper evaluates the performance of Naive Bayes and K-Nearest Neighbor models for predictive analysis. Using the Guns dataset, our results indicate that K-Nearest Neighbor outperforms Naive Bayes.*

I. INTRODUCTION

1.1 Data Mining

Data mining is the process of extracting valuable information and patterns from large datasets, also known as data discovery or pattern analysis. It aims to uncover hidden patterns that can inform decision-making and improve businesses. By analyzing past data and predicting future trends, data mining integrates statistics, AI, and database technology. Its applications are highly valued, driving increased sales and profitability for businesses. This predictive modeling is crucial across various fields, including engineering and medicine, where it forecasts outcomes and categorizes data for better understanding.

II. CLASSIFICATION

Classification is a widely used data mining technique, utilizing pre-classified examples to create a model for sorting records. Applications like fraud detection and credit risk assessment benefit from this analysis, often employing decision tree or neural network algorithms. The process involves learning and classification: analyzing training data to develop classification rules, and then using test data to validate these rules. Once validated, the rules can be applied to new data. For fraud detection, this entails compiling records of both fraudulent and legitimate activities for individual assessment. The classifier-training algorithm utilizes these predefined examples to determine the parameters necessary for accurate classification, encoding them into a model known as a classifier.

Kinds of grouping models:

- Classification by choice tree enlistment
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)

III. METHODOLOGY

This research work is focused on the performance evaluation of KNN and Naive Bayes Supervised learning.

3.1 Naive Bayes

The Naive Bayes is an enthusiastic method for game plan of quantifiable farsighted models. NB depends upon the Bayesian hypothesis [1]. This assessment utilizes class restrictive self-rule and has capacity to change rapidly. These depiction strategy evaluations the relationship between each property and the class for each manual for choose a restrictive likelihood for the relationship between the brand name attributes and the class [2]. Amidst setting up, the likelihood of each class is selected by checking how regularly it happens in the arranging dataset. This is known as the earlier likelihood; P(C=c). Regardless of the past likelihood, the calculation likewise enrols the likelihood for the occasion x given c with the supposition that the attributes are self-overseeing. This likelihood changes into the result of the probabilities of each single quality. The probabilities would then have the alternative to be assessed from the frequencies of the occasions in the arranging set.

3.2 K nearest Neighbour (KNN)

KNN is a highly effective algorithm with excellent performance and minimal training time. It's a simple yet powerful classifier, grouping samples based on their closest neighbors. As a non-parametric method, KNN retrieves the K nearest neighbors for a

given data record, forming a neighborhood. Typically, majority voting determines the classification, with distance-based weighting optionally considered. However, choosing an appropriate value for K is crucial for KNN's success, often requiring experimentation with different values to optimize performance.

IV. EXPLORATORY RESULTS

This segment portrays the exploratory outcomes acquired by applying the proposed grouping calculation to Guns dataset are taken from the UCI AI store [6]. In the Guns dataset, there are 1173 records, 14 credits and 2 class marks and the definite factual rundown are displayed in the figure-1 and figure-2. We have utilized the Python Language to analyze our proposed calculations. The Python Scikit-learn is a bundle for information order, relapse, grouping and perception. The order models were carried out in Python programming language.

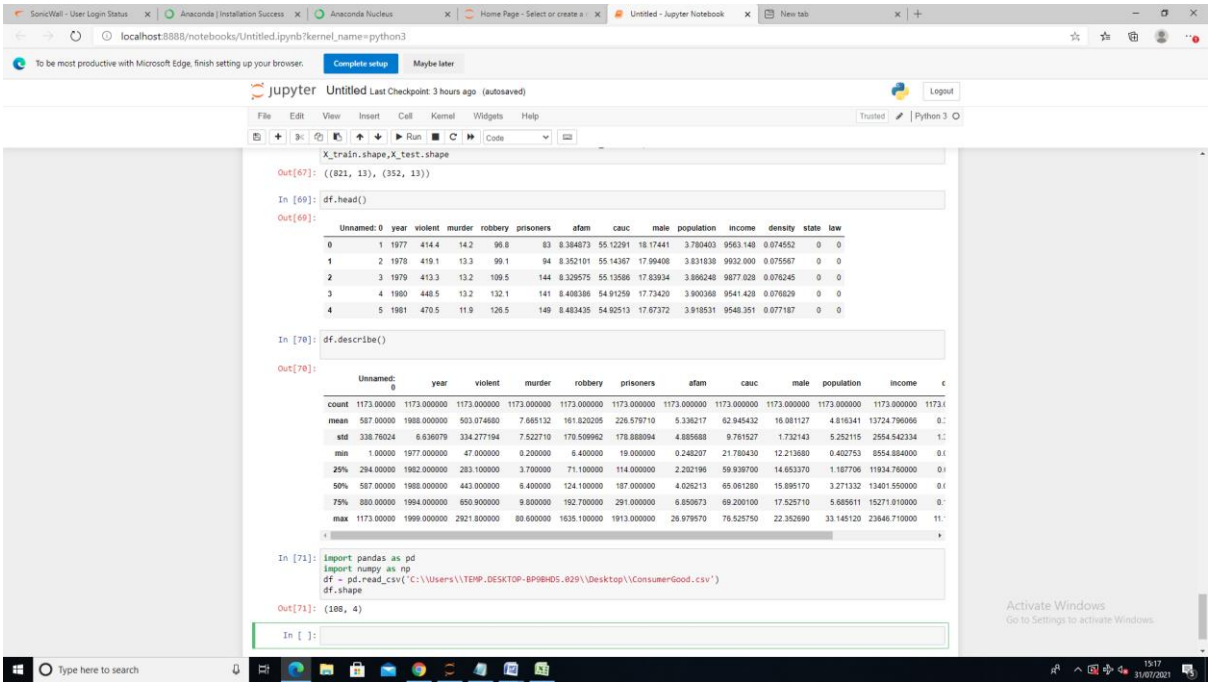

Figure-1: Statistical Summary of the Guns dataset

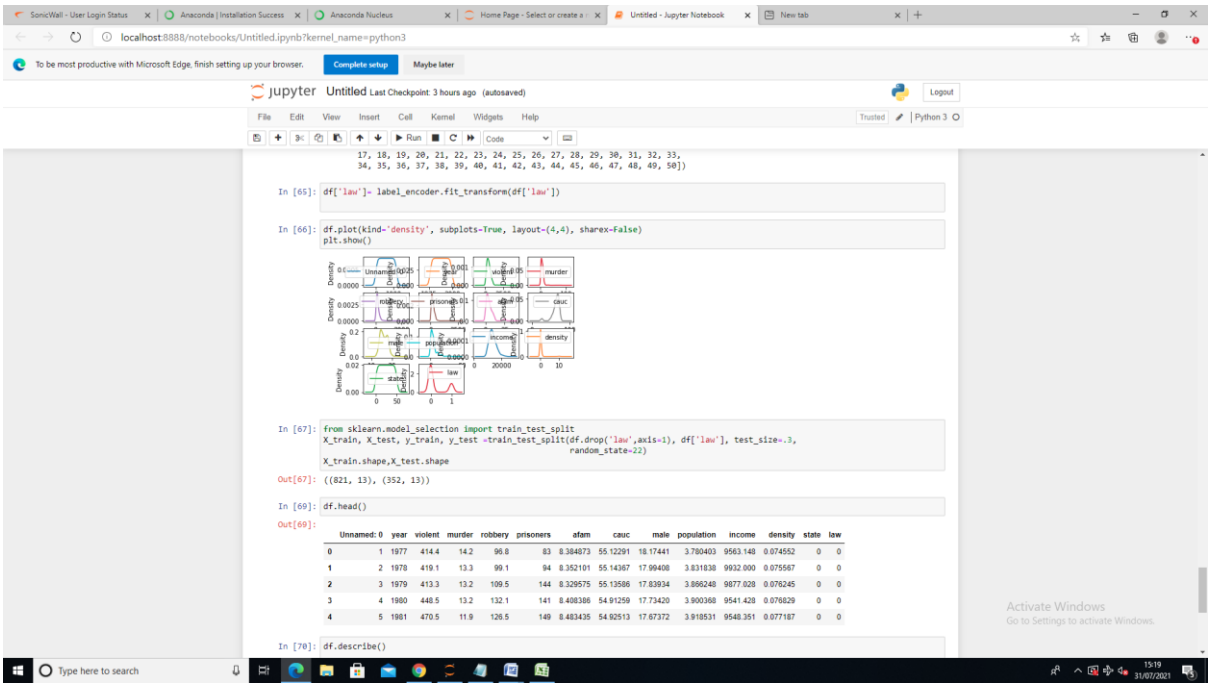

Figure-2: Density plot of each attribute details for the Guns dataset

4.1 Results

We utilize 70% of records as the preparation information and the other 30% as the testing information. The results of naïve bayes and multinomial naïve bayes classifiers are compared the on basis of correctly classified instances is shown in the table-1 and same shown in the figure-3 with their corresponding values.

Table-1
Performance of classifiers

Algorithm	Accuracy	Precision	Recall
Naive Bayes	90.45	91	91
KNN	91.76	92	92

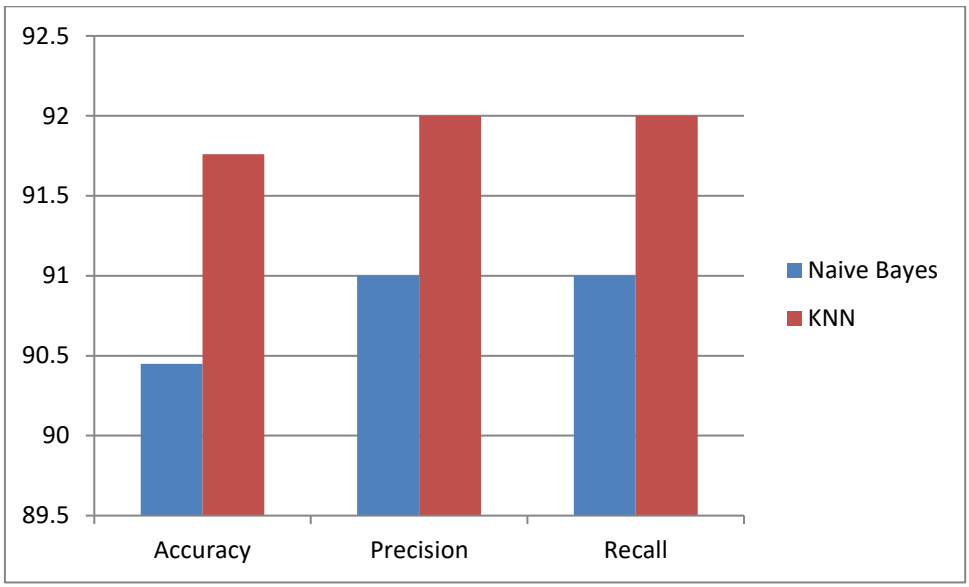

Figure-3: Performance of Classifiers

From the figure-3, we notice the exhibition of classification for Naive Bayes 90.45% of accuracy and the KNN has achieved the accuracy of 91.76%. So, the KNN classification has got highest accuracy when compared to Naive Bayes. So, the both algorithms have got highest accuracy, but only 1.31% difference of KNN when compared to Naive Bayes.

V. CONCLUSION

The KNN and Naïve Bayes approaches in AI offer theoretical and practical advantages. This paper explores their application in classifying the Guns dataset. Our findings indicate that KNN achieved higher accuracy compared to Naïve Bayes. We primarily focused on assessing the accuracy and performance enhancement of KNN.

REFERENCES

[1] Arthur Chol, NazgolTavabi, Adnan Darwiche, ―Structured Features in Naive Bayes Classification‖, Association for the Advancement of Artificial Intelligence, 2016.

[2] N. Friedman, D. Geiger, and Goldszmidt M. Bayesian network classifiers. Machine Learning, 29:131–163, 1997.

[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 Kaufman Series in Data Management Systems, 2nd Edition, 2006

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

[6] UCI Machine Learning Repository. http://archive.ics.uci.edu/ml