A Comparative Analysis of Naive Bayes and Logistic Regression Algorithms for Soybean Prediction

D. Thimmaraju

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

Abstract— *Soybean prediction plays a crucial role in optimizing agricultural practices, enhancing crop yield, and ensuring food security. Machine learning algorithms have proven to be effective tools in predicting crop outcomes based on various environmental factors. This research paper presents a comprehensive comparative analysis of two popular algorithms, Naive Bayes and Logistic Regression, for soybean prediction. The study aims to evaluate the performance of these algorithms in terms of accuracy, precision, recall, and F1-score. Additionally, we investigate the interpretability and computational efficiency of each algorithm to provide valuable insights for agricultural decision-making. The experimental results demonstrate the strengths and weaknesses of both algorithms and provide recommendations for selecting the most suitable algorithm for soybean prediction.*

I. INTRODUCTION

One of the most significant crops in the world is the soybean crop. A good source of protein for the human diet, in addition to being an important oil seed crop and livestock feed, is the importance of this crop. Since a decade ago, the demand for soybeans has grown, placing pressure on the supply. It's critical to boost crop yields in order to meet demand [5]. Four decades ago, the soy crop in India was first exploited, and ever since then, both its production and demand have skyrocketed. About 10% of all agricultural trade worldwide is made up of soybeans and their derivatives. The interest for Soybean and its items has quickly expanded since 1990s and has crossed the exchange for wheat and other coarse grains [6]. Be that as it may, different variables might meaningfully affect soybean crop development rate. These variables incorporate month, precipitation, temperature, hail, germination, seed, seed-size, leaves and so forth. Any abnormalities identified in any of these properties might defer plant development. Consequently, expulsion of such abnormalities becomes significant.

II. CLASSIFICATION

Grouping is perhaps of the most explored question in AI and information mining. In AI, characterization alludes to an algorithmic cycle for assigning a given info information into one among the various classifications given. A large number of genuine issues have been expressed as Order Issues, for instance credit scoring, liquidation expectation, clinical finding, design acknowledgment, text classification and some more. A calculation that executes grouping is known as a classifier. The information can be named as an occasion and the classifications are known as classes [2][7]. The qualities of the occasion can be portrayed by a vector of highlights. These highlights can be ostensible, ordinal, whole number esteemed or genuine esteemed. Order is a regulated methodology that figures out how to characterize new occasions in view of the information gained from a formerly grouped preparing set of examples.

III. METHODOLOGY

Naïve Bayes (NB) and Logistic Regression (LR) are two well known classifiers in AI [3]. As a matter of fact, NB is one of the quickest generative learning classifier for huge scope expectation and grouping errands on perplexing and inadequate datasets and simultaneously it can deal with both unmitigated (discrete) and non-all out (ceaseless) information. Then again, LR is typically a discriminative learning classifier [4] which is unmistakably produced for two class issues and can likewise deal with both unmitigated and non-straight out information. However, as a general rule, LR is a discriminative classifier, it very well may be worked as a generative classifier utilizing generative learning procedures. In this review, we assessed the presentation of the Credulous Bayes calculation on the Soyabeen dataset and contrasted it and the Strategic Relapse calculation.

3.1 Naïve Bayes' Classifier

Naïve Bayes is a probabilistic order technique that utilizes bayes hypothesis. It is "credulous" as in a quality worth on a given class is thought to be free of the upsides of different properties. The credulous bayes classifier takes a bunch of highlights from a dataset and decides the likelihood of each component happening in each class inside the information [1][3]. For each line of information, the upsides of the traits are utilized to ascertain the back likelihood for each class inside the dataset, the column of information is then allocated to the class with the most noteworthy back likelihood. This strategy is alluded to as credulous in light of the fact that it accepts that all elements of the dataset are autonomous of each other, which is a presumption that is logical false and subsequently guileless. In spite of this presumption not being valid in all cases, credulous bayes has been demonstrated to be a fruitful classifier in huge datasets.

Let X = (X1, X2, . . . , Xn) be an irregular variable and A1, A2, . . . , A be the properties of X related with the n parts X1, X2, . . . Xn separately (find in Figure 2). Let T = {x = (X1 = x1, X2 =x2, . . . , Xn = xn)} be the arrangement of preparing tests drawn from the number of inhabitants in X. Allow us to accept that there are c classes, C = {y1, y2, . . . , yc} and every single examples having a specific class marks Y = yj ∈ C. The undertaking of the classifier is to foresee the class name Y for a given example x. To anticipate the class mark of x, the credulous Bayes works out P(Y = yj|x) for each class yj , j = 1, 2, . . . , c and the example x is arranged in that class whose likelihood shows the most elevated esteem.

3.2 Logistic Regression

Logistic Regression is a notable method that efficiently utilized for displaying straight out results as an element of both persistent and clear cut factors in different applications. It is usually utilized for anticipating the likelihood of event of an occasion, in light of a few indicator factors that may either be mathematical or downright [2][4]. Allow us to think about the elements of the structure Y=f(X) or f:X→Y or P(Y|X) for the situation where Y is discrete-esteemed, and X= (X1, X2,...,Xn) is any vector containing discrete or constant irregular factors. Calculated relapse is one of the grouping calculations in AI for all out values like Yes or No, Valid or Misleading, 0 or 1. In this depiction, we consider the case just where Y is a boolean variable (say, either 0 or1), to work on documentation. Be that as it may, overall Y can be any finite number of discrete qualities.

IV. EXPLORATORY OUTCOMES AND INVESTIGATION

The examinations have been facilitated by utilizing Python programming vernacular. The Python Scikit-learn is a pack for information depiction, social occasion and depiction. We have considered the soybean data from the UCI man-made intelligence Vault [8] dataset for experimentation. The soybean dataset acquired from UCI AI store, contains 307 perceptions from soybean plants contaminated with 19 unique illnesses and 35 unmitigated qualities. The class wise imprint cases are shown in the figure-1.

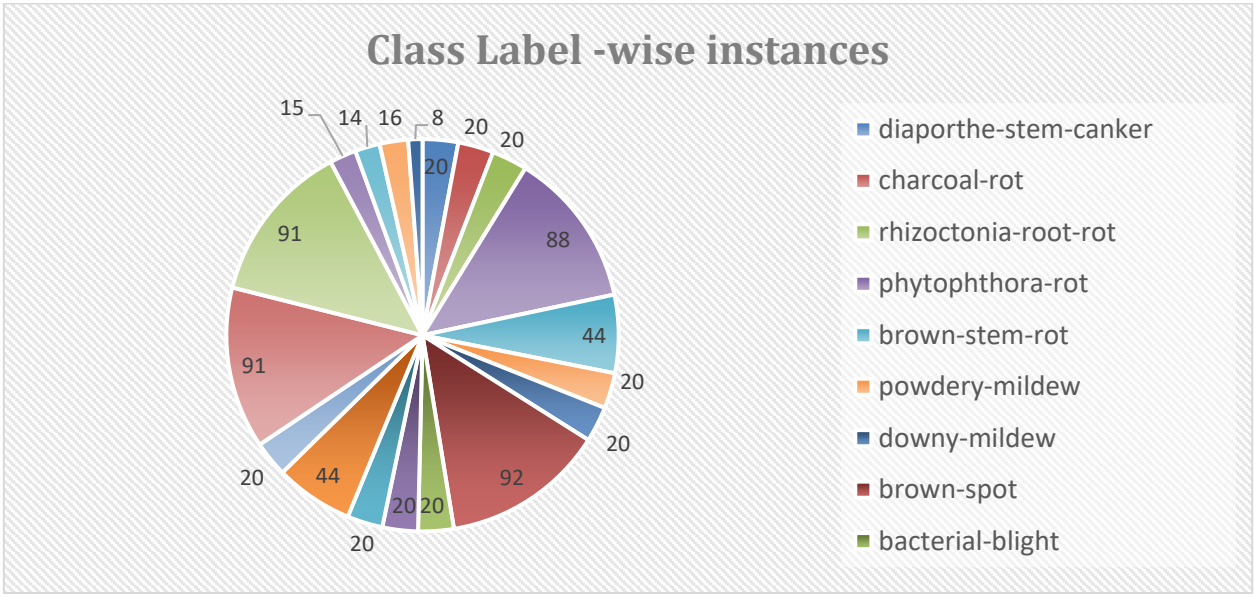

Figure-1: Class Label wise information

The entire dataset is partitioned for preparing the models and test them by the proportion of 70:30% separately. The preparation set is utilized to appraise each model boundaries, while the test set is utilized to survey the singular models freely.

4.1 Performance Evaluation

The proposed method uses different methodologies to improve the accuracy of the Soybean prediction. The proposed method uses the metrics such as accuracy, precision, recall and f-measure. It is proposed to classify a sample as positive when it is correctly classified TP (true positive) and FN (false negative) if it is classified as negative. TN (True Negative) means that a negative sample is classified as negative while the opposite is FP (False Positive).

Precision: The precision is calculated using the correctly classified positive crops to the total positives in the dataset. The below equation is used:

$Precision$ = TP/FP + TP

Recall: The correctly classified crops to total count of positive samples are calculated by the equation:

$Recall$ = TP/TP + FP

Accuracy: It is calculated by the ratio between correctly classified crop samples to total crop samples and the following equation is used for calculation:

$Accuracy$ = TP + TN / TP + TN + FP + FN

F-measure: $F1$-score is also known as F-measure. It is the harmonic mean of precision and recall is calculated by the equation:

$F-measure$= 2*Recall*Precision/ (Recall+Precision)

4.2 Results and Discussions

The presentation of the classifiers is assessed utilizing the generally utilized confusion matrix -based measurements, in particular, exactness, accuracy, and review. We study our two models utilizing arranged execution assessments like Exactness, Accuracy and Review, the Trial results are appeared in the table-1 and figure-2.

The figure-2 presents the performance metrics of two popular algorithms, Naive Bayes and Logistic Regression, for soybean prediction. The evaluation metrics include accuracy, precision, recall, and F1-measure.

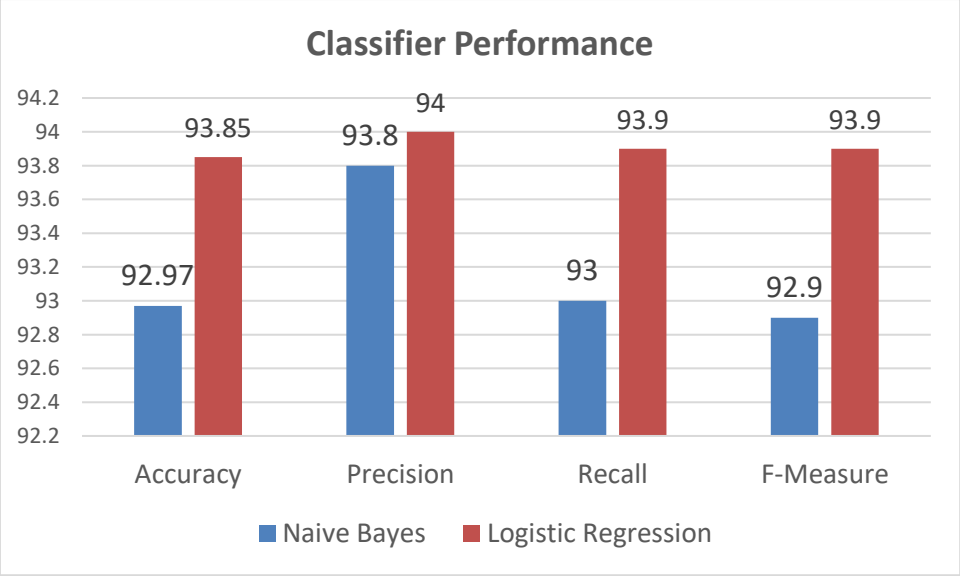

Figure-2: Performance of Algorithms

According to the results shown in the figure-2, the Naive Bayes algorithm achieved an accuracy of 92.97%. It demonstrated a precision of 93.8%, recall of 93.0%, and an F1-measure of 92.9%. On the other hand, Logistic Regression outperformed Naive Bayes with an accuracy of 93.85%. It achieved a precision of 94%, recall of 93.9%, and an F1-measure of 93.9%.

The performance of both algorithms indicates their effectiveness in predicting soybean outcomes. However, Logistic Regression exhibits slightly higher accuracy and precision compared to Naive Bayes. These results suggest that Logistic Regression may be better suited for soybean prediction tasks, providing more accurate predictions and minimizing the rate of false positives.

The higher precision values obtained by both algorithms indicate their ability to correctly identify positive instances of soybean prediction. A higher precision score implies a lower rate of false positives, which is crucial in agriculture as it reduces unnecessary costs and resources associated with false predictions.

Similarly, the recall scores for both algorithms are relatively high, indicating their ability to identify a large proportion of actual positive instances of soybean prediction. A higher recall score is beneficial for capturing a larger portion of positive instances, ensuring that potential issues or opportunities related to soybean crops are not overlooked.

The F1-measure, which considers both precision and recall, provides an overall assessment of the algorithms' performance. Both Naive Bayes and Logistic Regression achieved comparable F1-measures, indicating their ability to balance precision and recall effectively.

Although the difference in performance between Naive Bayes and Logistic Regression is relatively small, it is important to consider other factors such as interpretability and computational efficiency. Logistic Regression is known for its interpretability as it provides coefficients that can be analyzed to understand the impact of different features on the soybean prediction. On the other hand, Naive Bayes is computationally efficient and can handle large datasets efficiently.

V. CONCLUSION

In conclusion, both Naive Bayes and Logistic Regression algorithms exhibit promising performance for soybean prediction. While Logistic Regression demonstrates slightly higher accuracy and precision, Naive Bayes offers advantages in terms of computational efficiency. Therefore, the choice between the two algorithms should be based on the specific requirements and constraints of the soybean prediction task, considering factors such as interpretability, computational efficiency, and the importance of precision versus recall in the context of the agricultural decision-making process.

Further research can explore ensemble methods or other advanced machine learning techniques to improve soybean prediction accuracy and address the limitations of individual algorithms. Additionally, incorporating domain-specific knowledge and environmental factors into the prediction models can enhance their predictive capabilities and provide more valuable insights for soybean farmers and agricultural stakeholders.

REFERENCES

[1] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN:0973-8975.

[2] Dr. G. Ravi Kumar, K. Tirupathaiah and Prof. B. Krishna Reddy, “Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques”, International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019,

[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, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] Lee, T., Tran, A., Hansen, J., & Ash, M. (2016). Major factors affecting global soybean and products trade projections.

[6] Sharma, P., Dupare, B. U., & Pate, R. M. (2016). Soybean improvement through research in India and socio-economic changes, Indian Council of Agricultural Research, pp.1-3.

[7] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, “A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

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