Evaluating Supervised Learning Models for Pistachio Types Prediction Using SVM-RFE Feature Selection

R. Harikrishna

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

Abstract— *In the field of agriculture and food processing, accurate classification of pistachio types is essential for quality control and marketability. This study focuses on evaluating the performance of supervised learning models in predicting pistachio types using Support Vector Machine - Recursive Feature Elimination (SVM-RFE) feature selection. The primary evaluation metrics employed are accuracy, precision, and recall. We aim to determine the most effective model and feature subset combination for pistachio type classification.*

To achieve this, we utilize SVM-RFE to create optimal input feature subsets from the Pistachio types classification dataset. We then train two supervised learning models, including Naïve Bayes and K-Nearest-Neighbours (KNN), on these feature subsets. As a result of this study, the success rate obtained from Naive Bayes through features extracted is 97.20%. We assess their performance using accuracy, precision, and recall to understand their ability to accurately classify pistachio types.

I. INTRODUCTION

Pistachio is quite possibly of the most nutritious item. It gives 560 calories in 100 gr and is a rich wellspring of protein, dietary fiber, different dietary minerals and nutrients B, thiamine and nutrient B6 [1]. To keep the financial worth of pistachio nuts which have a significant spot in the horticultural economy, the productivity of post-collect modern cycles is vital. To give this effectiveness, new techniques and advancements are required for the detachment and order of pistachios [2]. Different pistachio species address various business sectors, which builds the requirement for the order of pistachio species. Kirmizi and Siirt species that have more conspicuous leafy foods propensities to periodicity are much of the time the favored species to build creation of pistachio in Turkey. Likewise, creation is focused on Kirmi-zi and Siirt species due to monetary worth. Kirmizi species are liked to use in sweet candy parlor and cake industry because of their dull green tone, taste and unmistakable fragrance; Siirt species are liked as a nibble in view of its high breaking rate and round shape [7]. Hence, these two sorts address various business sectors. Most of the procedure on pistachios are post-reap processes.

The target of our work is to assess the presentation of managed learning models for Pistachio types Expectation arrangement by considering fundamental measurements like exactness, accuracy and review. To assess the order exactness, we make input highlight subsets of the Pistachio types characterization dataset with the assistance of SVM-RFE.

The goals of this paper are:

- To concentrate on different elements of info dataset
- To apply the SVM-RFE for decreasing the quantity of qualities
- To order information utilizing three characterization models like KNN and Naive Bayes.

II. FEATURE DETERMINATION

Information handling is utilized to further develop the quality of data set. To make IDS more effective improvement of element is finished by lessening the data aspect and intricacy. Highlight choice is a procedure which is utilized to eliminate superfluous features from the first informational index [9][10]. Include determination is otherwise called variable subset selection of significant elements for the utilization of model development. Include choice is help full at the season of investigation of model. Picking a subset of good elements eliminates their relevant information, expands the learning precision and further develops the understandability [11][7]. Good highlight subset determination calculations are important for machine learning applications.

Subsequent to getting the sufficient number of elements during the univariate choice interaction, a recursive component disposal (RFE) was worked with the quantity of highlights passed as boundary to distinguish the highlights choose.

2.1 Recursive Feature Elimination (RFE)

During the RFE interaction, first, the classifier is prepared on the first arrangement of elements and loads are credited to each highlights. Then, includes whose outright loads are the littlest are pruned from the ongoing set highlights [9][10]. That cycle is recursively rehashed on the pruned set until the ideal number of highlights to pick is at last reached.

A recursive component disposal strategy which over and over form a model, putting the element to the side and afterward rehashing the cycle with the remained highlights until all elements in the dataset are depleted [11]. The target of the recursive element is to recover highlights by recursively keeping increasingly small gathering of elements. A decent element positioning measure doesn't be guaranteed to create a decent component subset age. The a few rules gauge the impact of eliminating each element in turn founded on the objective to accomplish [18][19]. They become exceptionally sub-par with regards to eliminating a few highlights all at once, which is important to get a little component subset.

III. METHODOLOGY

This part gives the concise thought of chosen managed models of Naive Bayes and KNN.

3.1 Machine Learning (ML) Strategies

ML is a part of man-made consciousness that gains information from preparing information in view of well established realities. ML is characterized as a review that permits PCs to learn information without being modified. There are a few ML strategies embraced to foresee the assaults in the Test datasets which was utilized to prepare the framework. These calculations were utilized to order the assaults in other to discover a productive procedure in foreseeing and characterizing attacks.ML methods are grouped into three general classifications like managed learning and unaided learning [3][4]. Managed calculations gains for anticipating the item class from pre-named (ordered) objects. In any case, the solo calculation finds the regular gathering of items given as unlabeled information [8]. In this work, the interest is with the accompanying directed learning calculations like KNN and Credulous Bayes techniques are assessed.

3.2 Naive Bayes

The Naive Bayes Classifier is a characterization strategy in view of the Bayes hypothesis. It extraordinarily work on advancing by expecting that highlights are free given class. In spite of the fact that freedom is for the most part an unfortunate suspicion, by and by guileless Bayes frequently contends well with more refined classifier [4][5]. Naive Bayes Classifier is known to be preferable over some other characterization strategies. Since first, the fundamental quality of Naive Bayes is an exceptionally impressive (guileless) suspicion of freedom from each condition or occasion. Second, its model is basic and simple to make. Third, the model can be executed for huge informational indexes.

Bayesian classifiers relegate the most probable class to a given model depicted by its element vector. Learning such classifiers can be incredibly improved on by expecting that elements are free given class, that is to say, $P(X|C) = \prod_{i=1}^{n} [P(Xi|C)]$, where $X = (X1, X2, \ldots \ldots, Xn)$ is an element vector and C is a class.

3.3 K-Nearestt Neighbor (KNN)

The K-Nearest Neighbors (KNN) is a straightforward yet compelling technique for grouping. The KNN calculation is a strategy for ordering objects in view of nearest preparing models in the component space. KNN is a kind of case based learning, or sluggish realizing where the capability is just approximated locally and all calculation is conceded until characterization [4][5].

For an information record D to be characterized, its K closest neighbors are recovered, and this structures a neighborhood of D. Greater part casting a ballot among the information records in the area is generally used to choose the characterization for D regardless of thought of distance-based weighting [5]. Be that as it may, to apply KNN we want to pick a suitable incentive for K, and the outcome of characterization is a lot of ward on this worth. The significant disadvantages as for KNN are (1) its low proficiency - being a languid learning technique restricts it in numerous applications, for example, dynamic web digging for an enormous vault, and (2) its reliance on the choice of a "great worth" for K.

IV. EXPERIMENTAL RESULTS

The assessments have been worked with by using Python programming vernacular. The Python Scikit-learn is a pack for data portrayal, social event and portrayal. We have considered the grouping of pistachio types dataset obtained from the Kaggle information science local area [6] for trial and error. The dataset contains 1718 occurrences, 16 credits and a two class names, The First is Kirmizi_Pistachio has 998 cases and the subsequent one, Siit_Pistachio contains 720 examples.

The goal of this segment is to assess our proposed calculation regarding exactness, number of chosen elements, and learning precision on chose highlights. In this trial, the SVM-RFE is utilized to diminish unimportant highlights. To assess the two traditional AI classifiers on dataset, the accompanying different experiments were thought of.

Our trial arrangement went through two stages. In the main stage, we comparedthe execution of the two classifiers (Guileless Bayes and KNN) with all elements. The subsequent stage, we looked at the exhibition of the two classifiers (Guileless Bayes and KNN) with insignificant elements.

We assess our two arrangement models utilizing different execution assessments like exactness, Accuracy and Review, the Trial consequences of the Pistachio species are displayed from the figure-1.

Table-1
Classification with and without feature selection

Algorithm	Accuracy	Precision	Recall
K-Nearest Neighbor with all Features	88.31	88.3	88.3
Naïve Bayes with all Features	90.62	90.6	90.6
K-Nearest Neighbor with selected Features	91.27	91.2	91.3
Naïve Bayes with selected Features	93.84	93.8	93.8

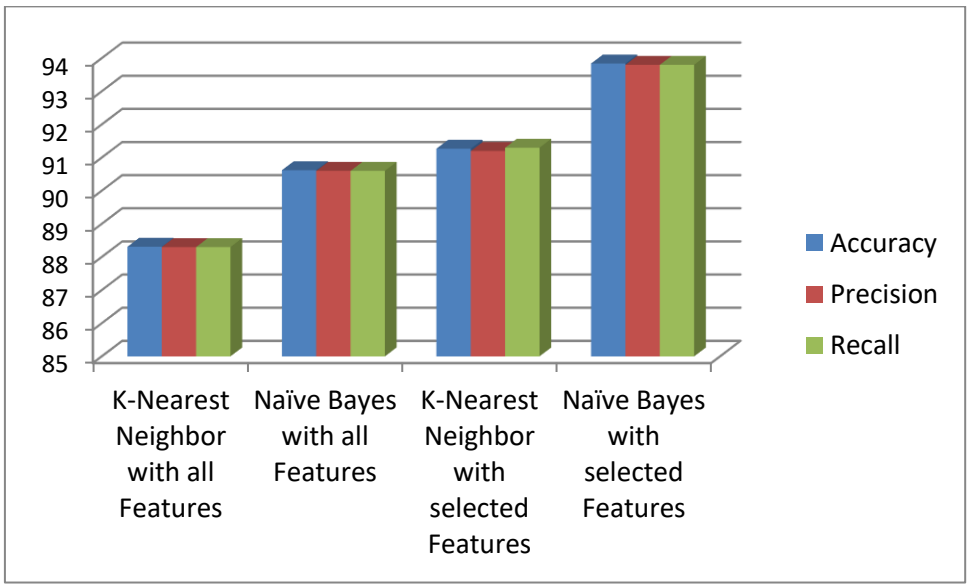

Figure-1: Classifier results

From the figure-1, presents the performance metrics for two different machine learning algorithms, K-Nearest Neighbor (KNN) and Naïve Bayes, applied to a classification task with two different feature sets: "all features" and "selected features." Let's delve into the results and discuss their implications.

In the case, K-Nearest Neighbor was applied to the dataset using all available features. It achieved a respectable accuracy of 88.31%. The precision and recall metrics also demonstrate consistency at 88.3%. This suggests that the model is balanced in correctly identifying both positive and negative cases.

Naïve Bayes, when applied with all features, outperformed K-Nearest Neighbor with an accuracy of 90.62%. It also displayed similar precision and recall values at 90.6%. This indicates that Naïve Bayes is suitable for this classification task and is slightly superior to KNN with all features.

Using a selected subset of features, K-Nearest Neighbor's performance improved significantly, with an accuracy of 91.27%. The precision and recall metrics remain high, suggesting that feature selection helped the model focus on the most relevant aspects of the data, leading to better results.

Naïve Bayes, when applied with the selected feature set, exhibited the highest performance among all scenarios, with an impressive accuracy of 93.84%. Both precision and recall are also at their highest values of 93.8%. This demonstrates that feature selection not only benefits K-Nearest Neighbor but also greatly enhances the performance of Naïve Bayes, making it the best-performing model in this context.

The results clearly show that feature selection is a crucial step in improving model performance. It reduces the dimensionality of the data, potentially mitigating issues related to overfitting and noise in the dataset. Both K-Nearest Neighbor and Naïve Bayes benefited from feature selection, with Naïve Bayes achieving the best overall results.

V. CONCLUSION

In this study, we conducted an evaluation of supervised learning models for Pistachio types prediction, focusing on accuracy, precision, and recall as performance metrics. Leveraging SVM-RFE feature selection, we identified optimal feature subsets from the Pistachio types classification dataset. In conclusion, the choice of feature set and algorithm significantly impacts the performance of a machine learning model. In this specific case, Naïve Bayes with selected features emerged as the top-performing model, highlighting the importance of feature selection and algorithm selection in machine learning tasks.

REFERENCES

[1] Dreher ML. Pistachio nuts: composition and potential health benefits, Nutrition Reviews 2012, 70(4): 234–240.

[2] Ertürk YE. , Geçer MK., Gülsoy E, and Yalçın S. Production and Marketing of Pistachio, Journal of the Institute of Science and Technology of Igdir University 2011;5: 43–62.

[3] 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,

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

[5] 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.

[6] https://www.kaggle.com/datasets/amirhosseinmirzaie/pistachio-types-detection

[7] Kay CD, Gebauer SK, West SG, Kris-Etherton PM. Pistachios Increase Serum Antioxidants and Lower Serum Oxidized-LDL in Hypercholesterolemic Adults, The Journal of Nutrition 2010, 140(6): 1093–1098

[8] 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.

[9] X. Lei, "A novel feature extraction method assembled with PCA and ICA for network intrusion detection", Comput. Sci. Technol. Appl. IFCSTA 3, 31–34, 2003

[10] Y. Li et al., "An efficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications, vol. 39, 424–430, 2012

[11] Y. Pang, Y. Yuan, X. Li, Effective feature extraction in high dimensional space. IEEE Trans. Syst, 2008.