1.

QUALITY PREDICTION OF WATERMELON USING RANKING FEATURE SELECTION METHODS AND MACHINE LEARNING ALGORITHMS

W. Pirunthavi[1], T. Sharnitha[1] and P. Mayuran[1]

[1]Department of ICT, Faculty of Technological Studies, University of Vavuniya

Abstract

This study was performed on the aim of detecting the quality of the watermelon with eight features; sound, color, root, belly button, texture, sugar rate, density, and touch which were obtained from the Kaggle website. Two ranking feature selection methods; ReliefF Ranking Filter and Information Gain Ranking Filter, and six machine learning algorithms; Decision Table (DT), J48 Tree (J48), Na ̈ıve Bayes (NB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest (RF) accordingly have been employed for the Feature Selection and Classification Model (FS-CM) to predict the quality of this fruit. Evaluation process has been conducted with five features which were selected under Information Gain Ranking filter. The metric Accuracy and ROC area were used for the evaluation and hence, MLP with IG was selected as the best model with the highest accuracy of 87.0813 detect the quality of the watermelon.

***Keywords**: classification model, information gain ranking filter, quality prediction, watermelon*

1 INTRODUCTION

The quality of the fruits or vegetables plays an important role in customer consumption and thereby affecting the sale of those fruits. The concept of quality is wide and covers several aspects such as external appearance, nutritional aspects, and presence of health-related compounds, safety and security [1]. The external appearance of the fruits is especially, the color and shape would make a prominent role in decision making. Further, the smell, texture and skin pattern make some impacts on the decision. However, the decision could vary according to the individuals with their human sensors. This manual prediction of the quality is time consuming and less effective and needs to learn from experience in choosing the quality and tasty fruits for their consumption [1]. The quality prediction-based researches were conducted on several fruits. These techniques could be applied to the watermelon fruit, in detecting the quality. Some methods include optical properties, sonic vibration [2], nuclear magnetic resonance (NMR), machine vision technique, electrical properties detection, computed tomography and electronic noses technique and so on [3].

2 METHODOLOGY

Determination of the quality of watermelon can be divided into five stages as shown below in Figure 1.

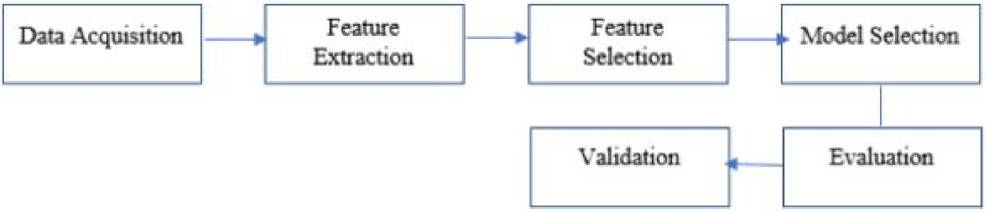

Figure 1. Methodology Diagram

2.1 Data Acquisition

The dataset was acquired from the Kaggle dataset [4].

Table 1. The Feature Set

Feature	Possible values	Nature of the feature
Color	Green, Dark green, light green	Nominal
Root	Rolled up, curly, straight	Nominal
Sound	Turbid, low, clear	Nominal
Texture	Clear, blurry, very blurry	Nominal
Belly button	Sunken, little sunken, flat	Nominal
Touch	Slippery, silky	Nominal
Density	0.233-0.779	Numeric
Sugar rate	0.038-0.469	Numeric

2.2 Feature Extraction

This dataset consists of 8 features. Features and the possible values are described in the Table 1 below.

2.3 Feature Selection

Above features are used for feature selection. Commonly two ranking algorithms were used for feature selection; ReliefF Ranking Filter, Information Gain Ranking Filter. These algorithms rank each feature in an order. In addition, both the orders are different from each other. Under each ranking method the first five features were selected for the analysis.

2.4 Model Creation

Some classification algorithms have been used for the quality detection of watermelon; such that, Decision Table, J48 Tree, Naïve Bayes, Support Vector Machine (SVM), MultiLayer Perceptron and Random Forest.

2.5 Evaluation

The metric such as Accuracy and ROC curve are used to evaluate and choose the best classification model that is suitable for our study. Three types of datasets were used in this study; full dataset (consisting of 8 features), 5 features from ReliefF Ranking filter, and 5 features from Information Gain Ranking filter.

2.6 Validation

The cross validation of 10 Stratified cross-validation were used for this study. This method is used to avoid the overfitting in the dataset.

3 Results and Discussions

Table 2. Feature Selection Methods

Feature Selection Methods	Ranked Attributes
ReliefF Ranking Filter	**Sound, Color, Root, Belly button, Texture**, Sugar rate, Density, Touch
Information Gain Ranking Filter	**Density, Sugar rate, Belly button, Texture, Root**, Sound, Color Touch

Table 3 illustrates the accuracy level of each selected classification algorithms.

Table 3. Accuracy of Six Classification Algorithms

Classification Algorithms	Full Dataset (F)	ReliefF (R)	Information Gain (IG)
Decision Table (DT)	86.6029%	84.2105%	85.6459%
J48 Tree (J48)	84.2105%	84.2105%	85.1675%
Naïve Bayes (NB)	75.5981%	76.555%	81.3397%
Support Vector Machine (SVM)	81.3397%	81.8182%	81.8182%
Multi-Layer Perceptron (MLP)	87.0813%	85.1675%	87.5598%
Random Forest (RF)	82.7751%	85.1675%	84.689%

The accuracy of the Multi-Layer Perceptron outperforms the other classification algorithm with the highest accuracy in three datasets. Comparing the three datasets, the MLP with the Information Gain ranking filter stands out with the accuracy of 87.5598%.

Table 4 describes the evaluation metric of TP Rate, FP Rate, Precision, Recall, F-Measure and MCC for 18 Feature Selection-Classification Models (FS-CM). ROC area was compared with the MLP classification for three datasets. Hence, Figure 2 shows the variation of ROC area against the dataset used.

MLP with IG shows the highest with 0.86 The confusion matrix for MLP-IG is shown in the Table 5.

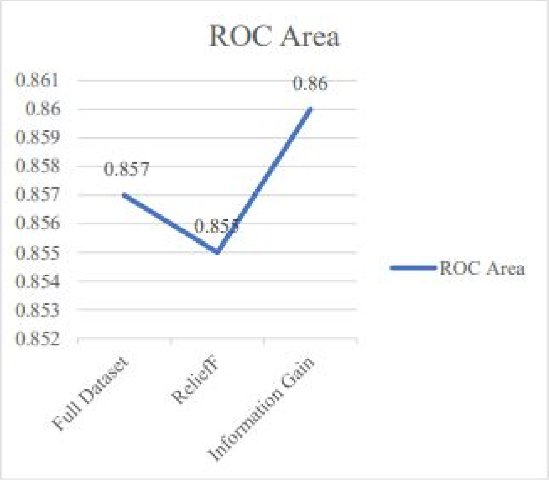

Figure 2. Major issues encountered by the tea processors

Table 4. Feature Selection-Classification Models (FS-CM)

FS-CM	TP	FP	Precision	Recall	F-Measure	MCC
F-MLP	0.871	0.163	0.870	0.871	0.870	0.722
F-SMO	0.813	0.267	0.820	0.813	0.805	0.598
F-NB	0.756	0.302	0.752	0.756	0.752	0.470
F-J48	0.842	0.195	0.841	0.842	0.841	0.660
F-DT	0.866	0.166	0.865	0.866	0.865	0.712
F-RF	0.828	0.204	0.827	0.828	0.827	0.631
R-MLP	0.852	0.199	0.853	0.852	0.849	0.680
R-SMO	0.818	0.264	0.827	0.818	0.810	0.610
R-NB	0.766	0.267	0.765	0.766	0.765	0.500
R-J48	0.842	0.215	0.844	0.842	0.838	0.660
R-DT	0.842	0.215	0.844	0.842	0.838	0.660
R-RF	0.852	0.199	0.853	0.852	0.849	0.680
IG-MLP	0.876	0.160	0.875	0.876	0.874	0.733
IG-SMO	0.818	0.264	0.827	0.818	0.810	0.610
IG-NB	0.813	0.267	0.820	0.813	0.805	0.598
IG-J48	0.852	0.184	0.851	0.852	0.850	0.681
IG-DT	0.856	0.172	0.856	0.856	0.856	0.692
IG-RF	0.847	0.177	0.846	0.847	0.846	0.67

4 Conclusion

This study was performed on the aim of detecting the quality of the watermelon with eight features, from the Kaggle dataset. Two ranking feature selection algorithm and six classification algorithms

Table 5. The confusion matrix for MLP-IG

Prediction	Quality	Not Quality
Quality	62	17
Not Quality	10	120

have been employed for the Feature Selection and Classification Model (FS-CM). Evaluation process has been conducted with five features selected under Information Gain Ranking filter. The metric Accuracy and ROC area were used for the evaluation and hence, MLP with IG was selected as the best model with the highest accuracy of 87.0813

ACKNOWLEDGMENT

We would like to express my very great appreciation to Ms. Rukshani Puvanendran for her valuable and constructive suggestions during the planning and development of this wonderful research on this topic. Further, We would also like to thank the staff of the Department of ICT, Faculty of Technological Studies, University of Vavuniya for providing this golden opportunity.

REFERENCES

[1] Wan Nurul Suraya Wan Nazulan, Ani Liza Asnawi, Huda Adibah Mohd Ramli, et al. "Detection of Sweetness Level for Fruits (Watermelon) With Machine Learning". In: *2020 IEEE Conference on Big Data and Analytics (ICBDA)*. IEEE. 2020, pp. 79–83.

[2] Chengqiao Ding, Zhe Feng, Dachen Wang, et al. "Acoustic vibration technology: Toward a promising fruit quality detection method". In: *Comprehensive Reviews in Food Science and Food Safety* 20.2 (2021), pp. 1655–1680.

[3] Haisheng Gao, Fengmei Zhu, and Jinxing Cai. "A review of non-destructive detection for fruit quality". In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2009, pp. 133–140.

[4] *Watermelon Quality*. 2022. URL: https://www.kaggle.com/datasets/chenshiji/watermelon-quality-prediction-data.