An Analytical Report for developing Computational Model for the Prediction of Yield and Strawberries Firmness based on Air Humidity

Abstract

High air humidity reduces strawberry production and shelf life. It is necessary to maintain the air humidity level throughout the whole time of cultivation. This study presents an artificial intelligence approach to model yield production and fruit firmness based on past data. This model will help farmers to take appropriate actions in time if current environmental factors are not favorable for high production and better fruit shelf life.

Contents

1 Introduction

The humidity level in the air affects Strawberry cultivation such as a high humanity level produces less production with poor fruit firmness. To get the maximum production and increase the shelf life of fruits, the farmers need a mechanism that provides them to anticipate their level of production and the fruit firmness level. The prediction will help farmers to take appropriate actions in time to get maximin yield production and have a better fruit shelf life.

In this study, we investigated the effect of two main factors, i.e. temperature and humidity level for getting maximum Strawberry production and better fruit firmness. The study is based on the collection of data samples which consists of the temperature and the corresponding level of humidity, infrared temperature, and dew point. The data also provided the value ranges for maximum production and better shelf life (see Table 1 and 2).

Table 1: Range of humidity to get maximum yield production

Air Humidity Range	Yield Production
65 to 75%	High

Table 2: Range of fruit firmness based on air humidity ranges

Air Humidity Range	Fruit Firmness Range
56 to 81%	3.2 to 9.9 N
60 to 86%	5.4 to 10.4 N

We developed a computational model which takes air temperature, humidity level, infrared temperature, and dew point as input and predicts two outcomes, i.e. whether the farmer will get maximum yield or not and the fruit firmness range. The proposed system is based on supervised learning (i.e, multilabel classification) which learns the hidden patterns with respect to yield production and fruit firmness.

Further in this report, Section 2 explains the procedure to get multilabel for each sample in the data and the model for classification, Section 3 presents the details and results of the experiments and lastly presents the conclusion in Section 4.

2 Theoretical Analysis

The first step in supervised learning is to get the labels for each sample available in the dataset. The information provided in Table 1 and Table 2 is not enough to get the unique label for each sample. We can see that classes are overlapping, i.e the humidity ranges for maximum yield production occupy two fruit firmness ranges. This type of problem is called mutually non-exclusive. This is the reason that we treated this problem as a multilabel classification problem.

We extracted an important formation from Table 1 and Table 2, i.e. air humidity has a linear relationship with fruit firmness. The available humidity range starts from 56% and ends at 86% having fruit firmness from 3.2 to 10.4 N respectively. Based on this information, we used linear interpolation to get intermediate ranges for humidity and firmness level. This will help us to detect the class boundaries in which each data sample lies. We categorized each data sample into 6 classes; 2 for yield production and 4 for firmness (see Table 3 and 4). Furthermore, Table 5 represents two annotated data samples as an example.

Table 3: Yield production classes based on air humidity ranges

Class Type	humidity Range	Class Description
A	65 to 75%	High Production
B	<65 and 75%>	Low Production

Table 4: Range of fruit firmness based on air humidity ranges

Class Category	humidity Range	Fruit Firmness
C	<56%	<3.2 N
D	56 to 65.99%	3.2 to 5.5 N
E	66 to 75.99%	3.6 to 7.9 N
F	76to 86.00%	8.0 to 10.4 N

For modeling, we choose logistic regression in a chain manner. In a chain manner, a single logistic regression model is trained for each class and each model also considers the prediction of the earlier model in the chain.

Table 5: Data features and annotated class labels

Airtemp	Airhumidity	Irtemp	Dewpoint	Class Label	One-hot Encoding
14.77	68.60	16.27	29.75	A and E	[1,0,0,0,1,0]
13.37	77.30	14.78	28.14	B and F	[0,1,0,0,0,1]

The dataset consists of other features such as device id and the time but we dropped these features in our modeling. The reason behind this is that we only considered the established correlation i.e. humidity with firmness and humidity with yield production. For the time feature, we are considering air temperature which alternatively represents the time. This alternative is given (see Table 6)

Table 6: Temperature with respect to time

Time	Temperature
Day	18 to 24 $^{\circ}C$
Night	10 to 13 $^{\circ}C$

2.1 Dataset Statistics

This section provides important statistics related to the dataset. The dataset has 16,105 samples in it and a few samples contain *nan* values which we replaced with the columns means. The total number of the missing sample is 110. The dataset is split into train and test sets and the split ratio is 70% and 30% respectively.

Fig. 1 represents the total number of samples per class. We can see clearly that class A and class E have the majority as compared to other classes. Similarly, Fig. 2 interprets that when the farmer has high production then the fruit firmness level is between 3.6 to 7.9 N (i.e. Class E). On the other hand, Fig. 3 shows that when the farmer has low yield production then the fruit firmness range is between 3.2 to 5.5 N (i.e. Class D).

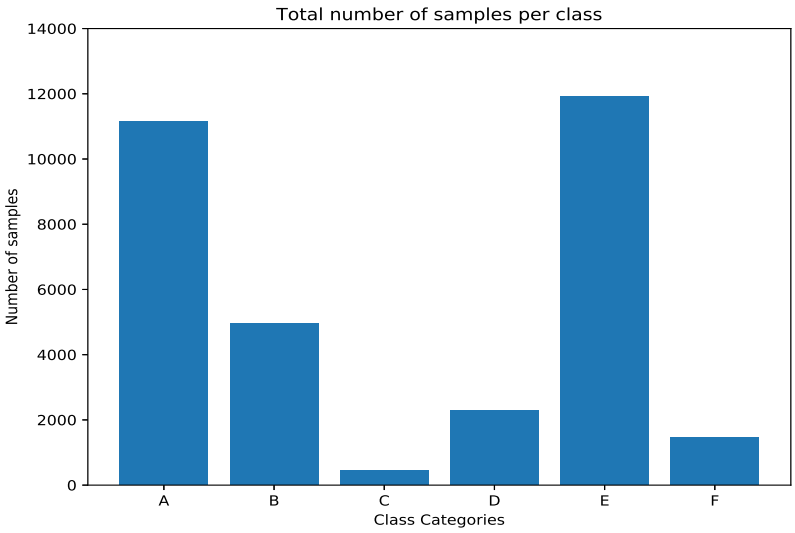

Figure 1: Total number of samples per class

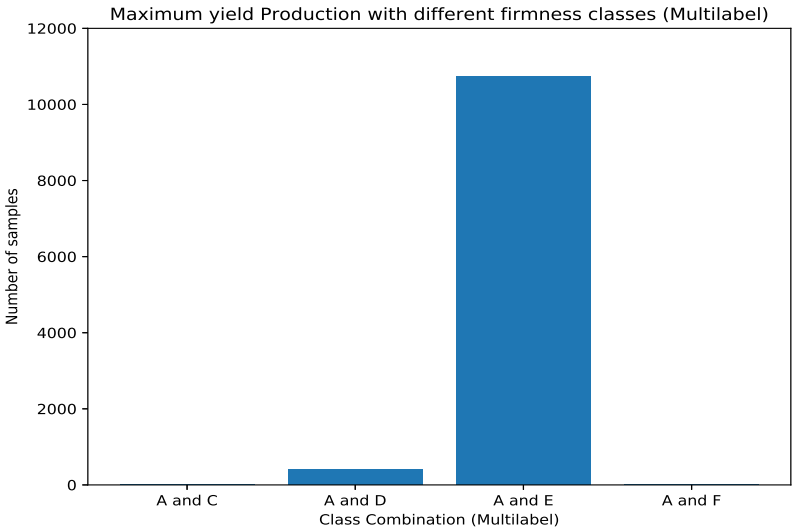

Figure 2: Number of samples for high production class with respect to each firmness class

3 Validating the Model

To validate the proposed approach, we tested using 5 cross-validation folds and evaluate the model performance based on the f1-score of each class (see Table 7). The results show that

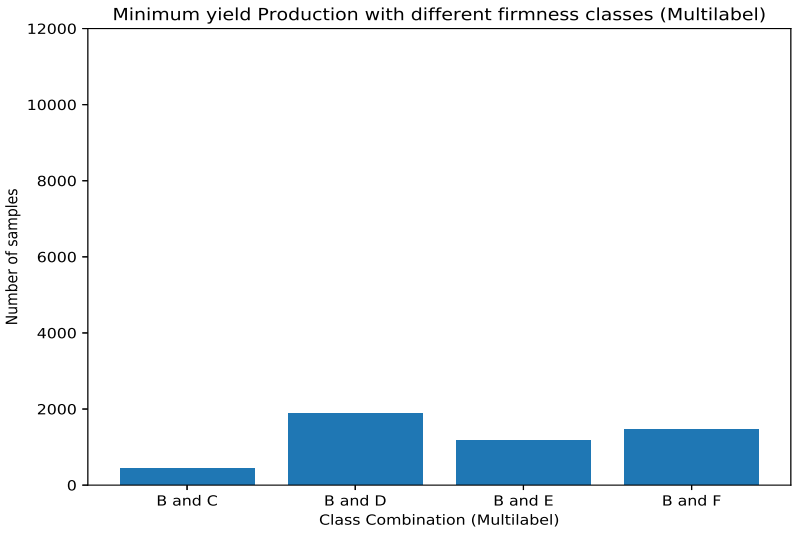

Figure 3: Number of samples for low production class with respect to each firmness class

the proposed model effectively models the relationship between environmental factors with yield production and the fruit firmness level.

Table 7: Model Performance

Class Category	F1-score
Class A	0.94
Class B	0.78
Class C	0.99
Class D	0.96
Class E	0.99
Class F	0.98

3.1 Qualitative Analysis

This section represents the qualitative analysis of the model. Table 8 presents a few testing samples and their interpreted corresponding results.

Table 8: Model predictions and their interpretations. Green represents true prediction and red represents the wrong prediction

Samples	Labels	Prediction	Interpretation
[22.69,58.57,21.97,23.77]	[0,1,0,1,0,0]	[0,1,0,1,0,0]	The farmer will have low productivity with fruit firmness range between 3.2 to 5.5 N
[14.61,68.28,15.87,29.63]	[1,0,0,0,1,0]	[1,0,0,0,1,0]	The farmer will have high productivity with fruit firmness range between 3.6 to 7.9 N
[23.14,58.54,22.47,22.21]	[0,1,0,1,0,0]	[0,1,0,1,0,0]	The farmer will have low productivity with fruit firmness range between 3.2 to 5.5 N
[14.14,77.67,15.47,27.40]	[1,0,0,0,0,1]	[1,0,0,0,0,1]	The farmer will have high productivity with fruit firmness range between 8.0 to 10.4 N
[31.50,50.97,3.13,15.60]	[0,1,1,0,0,0]	[1,0,1,0,0,0]	The farmer will have high productivity with fruit firmness range less than 3.2 N
[21.10,50.91,22.99,26.65]	[0,1,1,0,0,0]	[0,1,1,0,0,0]	The farmer will have low productivity with fruit firmness range less than 3.2 N
[23.96,55.16,23.69,22.33]	[0,1,1,0,0,0]	[1,0,0,0,1,0]	The farmer will have high productivity with fruit firmness range between 3.6 to 7.9 N

4 Conclusion

In this study, we proposed a machine learning approach that models the environmental factors to yield and fruit firmness prediction. The proposed approach is based on the chain of logistic regressions and the results show that the model has got a good performance in order to predict yield and fruit firmness. The classes present in the dataset are overlapping each other. Although, we tried our best to build a boundary between classes still some overlapping is present. This is responsible for the presence of noise in the dataset. To reduce the noise, more data is needed.