SUMMARY OF TOMATO DATA

SANDYA DE ALWIS, MYUNG NA, AND GANG LI

Contents

1.	Introduction	1
1.1.	Data sets	1
1.2.	Outside data	2
1.3.	Inside data	2
1.4.	Plant data:	2
2.	Preliminaries	5
3.	Problem Statement	7

1. Introduction

The tomato (Solanum lycopersicum) is a herbaceous plant with highly divided leaves that have long, lean hairs with a characteristic aroma. The flowers are bright yellow and have a bottle-shaped stamen cone in the center. The fruit is usually bright red, but can be many different colors. There are many varieties of the tomato. Originally, the tomato was from the western coastal deserts of South America, but today grows all over the world.

These tomato plants were cultivated under five different supervision in five different locations in South Korea. The aim of project is study on productivity improvement of crop yield. There are two growth habits in tomatoes namely determinate and indeterminate. In this experiment, indeterminate tomato plants were chosen. Tomato planted on the 26th week of the year(end of June) in plant houses with control environment. Recordings were taken trough out the year which is represent one life cycle of the plant.

1.1. **Data sets.**

There are three types of data sets. First type of data sets contain only environmental data. This environment data can be divided in to further two groups namely outside data and inside data. These eight data sets contain 9 attributes namely circulation temperature, inner temperature, heat temperature, outside temperature, inner humidity, radiation, cumulative radiation, inner temp1 and inner temp2. Data were collected using sensors. Each attributes contain data for every minuet for one month except March 2015 .(It's contain only 13 days).

Biomasdata(bea) set hold 12 attributes with 4 replicates over 52 weeks. These attributes are Length of Grow, Length of Leaves, Width of leaves, Number of Leaves, Stem-Diameter, Flower from Top, Flower position, Set position, Harvest position, Number of fruits, Production(kg) and average Weight of Fruit. Other five

1991 *Mathematics Subject Classification.* Artificial Intelligence.

Key words and phrases. Machine Learning, Data Mining, Tomato...

data sets hold combination of environment and biomass data. Means of all the raw data per week and other more attributes represent in this set. Each set contain 49 columns 52 rows and around 2500 records. Table 1 summarise all the data sets.

Table 1. Tomato Data Set

Data set Number	Name	Number of Attributes	Number of Rows	number of records
1	bae-ev-2014-08	9	44609	401,481
2	bae-ev-2014-09	9	43110	387,990
3	bae-ev-2014-10	9	44638	401,742
4	bae-ev-2014-11	9	43173	388,557
5	bae-ev-2014-12	9	44626	401,634
6	bae-ev-2015-01	9	44638	401,742
7	bae-ev-2015-02	9	40308	362,772
8	bae-ev-2015-03	9	18229	164,061
9	biomassdata(bea)	12	52 with 4 replicates	2496
10	bea14-15	49	51	2499
11	gu14-15	49	52	2548
12	moon14-15	49	52	2548
13	shin14-15	49	52	2548
14	sung15-16	49	52	2548

1.2. **Outside data.** In summarized data sets, first three attributes represent outside temperature. They are average temperature for 24 hours, maximum ambient temperature and minimum ambient temperature. Temperature is one of the most important factor for tomato plantation. According to [**?**] temperature directly effects the harvest. Solar Radiation during the twenty four hours is the next out side environment factor. This is significantly and positivity correlated to yield specially during the days before anthesis [**?**].

1.3. **Inside data.** Recording of environment data inside the plant house can be categorised into six main factors. They are temperature, humidity, CO2 level, water, EC and pH. As explained, temperature is the effect for tomato yield, especially day time. At night time, humidity also effect the quality of fruit. Therefore, humidity is also an important factor. CO2 level day(ppm) is one of the key elements in photosynthesis. When considered about photosynthesis, water is another key element. There are many attributes explaining about water such as water supply, intake and drainage.

Gift EC(Electric conductivity?) [this is not clear (G-EC Supply Average)] In hydroponics cultivations, it is essential maintain correct pH level all the time. Here, Trans1 means amount of radiation received to the green house. This amount depend on the construction material and number of layers of the plant house.

1.4. **Plant data:** Several factors of the plants were recorded trough out the year. No data for some attributes during first few weeks. For an example no harvest data during first few weeks. Different growth stages of the tomato plant shown in the figure 1.

Growth can be measured using different parameters. Plant height is basic measurement and in this data set it is mentioned as growth length. Cumulative growth is a term used to describe a percentage of increase over a set period of time.

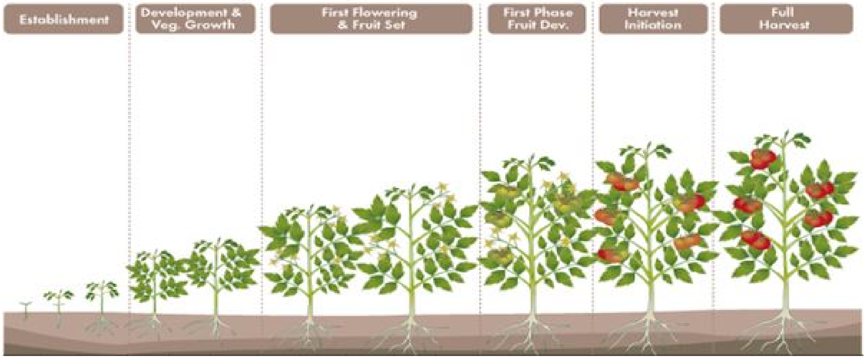

FIGURE 1. Growth stages of Tomato plant

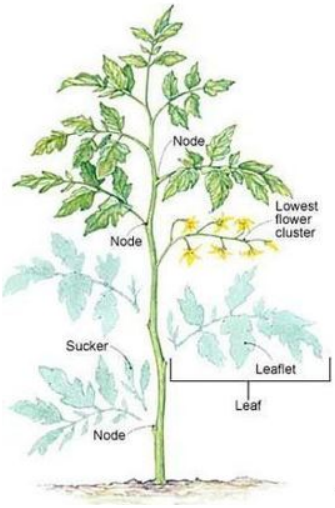

FIGURE 2. Plants parts

Tomato plants part are illustrated in figure 2. As in the picture, leaves emerge through the nodes. Sometimes suckers also emerge through the nodes. However, in tomato cultivation it is advised to remove suckers to maintain the plant. Flower clusters start from the lower part of the stem and continue upwards.

The way of measuring leaf length and leaf width of tomato is shown in figure 3. However, here length and width of leaflet were measured. Number of leaves per plant is directly related to production because photosynthesis happening in the leaves.

When measured, the thickness of the stem is important to keep in mind to measure in the same height in all the plants because stem thickness is not the same from top to bottom. In this data set stem thickness was taken between third and fourth nodes. Height of the flower is another attributed taken in this data set.

Days of yield is another reading in this data set. However, number of fruits and number of fruits per unit are more important factors especially in crop like tomato. Flowering speed and fruit set speed are also recorded here. Leaf Area Index(LAI)

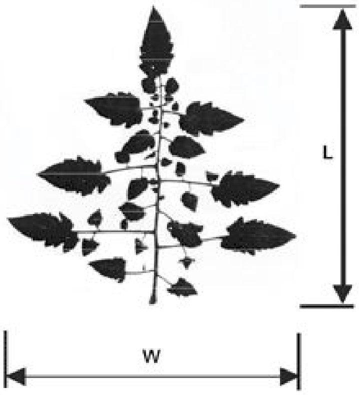

Figure 3. Leaf measurements of Tomato

is a key factor in this data set. It is a main feature in agricultural industry and plant physiology. In this data set, the next attributes is factor of fruit, which is consider as weight of tomato per standard weight of tomato. Here, standard weight consider as 175g. The attribute trans1 means amount of solar radiation received to the pant house. (Transmittance = 0.85: This value depend on the state of green house). Trans2 means amount of radiation received from the plant and PED is another attribute of plant factors and the last attribute is average weight per unit.

The rest of this report reviews all the attributes summarized in the table 2 and 3. After that results of preliminaries were discussed. Finally, current difficulties and problems that still exist in increasing yield were discussed.

Table 2. Tomato attributes summary

Attribute No	Attribute name	Description	Data type
1	ID		
2	week	week of the year	Time
3	y	Yield	plant Factor
4	x1	average temperature for 24hours	Outside environment
5	x2	maximum ambient temperature	Outside environment
6	x3	minimum ambient temperature	Outside environment
7	x4	24h Radiation sum(J)	Outside environment
8	x5	Average temperature for 24 hours	Inside environment
9	x6	average temperature(day)	Inside environment
10	x7	average temperature(night)	Inside environment
11	x8	average humidity(day)	Inside environment
12	x9	average humidity (night)	Inside environment
13	x10	average maximum humidity	Inside environment
14	x11	average minimum humidity	Inside environment
15	x12	average CO2 level day(ppm)	Inside environment
16	x13	Water (gift-driper)	water supply
17	x14	Water (gift-no)	water supply
18	x15	Water (gift)	water supply
19	x16	Water (gift)	water supply
20	x17	Water (drain)/slab	water supply
21	x18	Water (drain)	water supply
22	x19	Water (cc/J)/	water supply
23	x20	Water uptake/	water supply
24	x21	Water drain (cc/J)/	water supply
25	x22	Water (drain/gift)	water supply
26	x23	Gift EC	water supply
27	x24	Gift pH	water supply
28	x25	Slab EC	water supply
29	x26	Slab pH	water supply
30	x27	growth length	Plant factors
31	x28	cumulative growth	Plant factors
32	x29	length of leaf	Plant factors
33	x30	width of leaf	Plant factors
34	x31	no of leaves	Plant factors
35	x32	Thickness of stem	Plant factors
36	x33	height of flower	Plant factors

2. Preliminaries

Preliminary test were conducted to identify more details about these summarized data sets. Firstly, percentages of null or missing values were calculated individually and as one set. All the results were tabulated in table . When compared, all the attributes, most of the plant factors data are not available. In the beginning of the experiment most of the reading, such as number of fruits, yield are not available to measure may be the reason. According to results least percentage of missing or null shows in 'moon$14-15$'data set. On the other hand, highest percentage of missing or null showed in gu$14-15$ and shin $14-15$ respectively.

Table 5represents Data type, mean, standard deviation and if more clarification is needed or not respectively. Data can be categorised basically into two groups i.e. Categorical data and continuous data. Here we tried to categorize all the attributes. (However, some attributes are doubtful). Mean and the standard deviation of all the attributes were calculated and presented in next two columns of table 4.

3. Problem Statement

In order to predict harvest, it is essential to handle and manage all data sets of the parameters measured. Considering the amount of data available and to distinguish the pattern and extent of relationships for useful and efficient extraction of knowledge, there is a need for data mining techniques.

When considering current research, such as [**?**], [**?**] and [**?**] yield, number of fruits and weight of fruits can be predicted from solar radiation, temperature, water uptake etc. As mentioned in [**?**] solar radiation with the number of days can be predicted by number of fruits per plant via anthesis.

The primary objective of the research is to offer scientific analysis of data. The analysis is to be done by applying machine learning and data mining techniques to the data sets. Some Initial works were done on the data set with the use of simple analytical techniques. The main problem is which and when environment factors effect the yield? In conclusion, some of the research question are:

- Which week/s environment factors affect to the yield?
- How can we identify the most prominent factors that affect the crops at different stages of growth ie. each week?
- Which function can efficiently describe yield pattern in data set?
- What is the relationship between yield and attributes such as temperature, humidity, solar radiation etc. ?
- What is the best machine learning algorithm that gives best prediction?
- What are the best strategies to maximize the overall yield?

(A. 1) School of Information Technology, Deakin University, 221 Burwood Highway, Vic 3125, Australia
Email address, A. 1: sdealwi@deakin.edu.au

(A. 2) Data provider

(A. 3) School of Information Technology, Deakin University, 221 Burwood Highway, Vic 3125, Australia
Email address, A. 2: gang.li@deakin.edu.au

Attribute No	Attribute name	Description	Data type
37	x34*		Plant factors
38	x35*		Plant factors
39	x36*		Plant factors
40	x37	Days of yield	Plant factors
41	x38	no of fruits	Plant factors
42	x39	no of fruit per unit	Plant factors
43	x40	Flowering speed	Plant factors
44	x41	set speed	Plant factors
45	x42	$LAI = \frac{(Leaflength)*(Leafwidth)*(0.5)*(Leafnumber)*(DENSITY)}{(10000)+(Penetrationarea)}$	Plant factors
46	x43	$Factoroffruit = \frac{weightoftomato}{standardweightoftomato}$	Plant factors
47	x44	$Trans1 = 24hrRadiationsum(J) * Transmittance : (Transmittance = 0.85)$	Environment factors
48	x45	$Trans2 = (LAI)^3 - (0.133 * Leafarea(LAI)^2) + (0.606 * Leafarea(LAI) + 0.003)$	Plant factors
49	x46	$PED = \frac{((Numberoffruits/)}{DENSITY} * Factor * referencelight + basicmetabolism * 7$	Plant factors
50	x47	average weight per unit	Plant factors

Table 4. Null or Missing Data Percentage

Attribute No	Attribute name	all	bea14-15	gu14-15	moon14-15	shin14-15	sung15-16
y	Yield	25	24	26	10	69	31
x1	avg temp	9	12	21	12	20	8
x2	max temp	9	12	21	10	22	8
x3	min temp	11	12	21	12	20	6
x4	Radiation(J)	11	12	21	10	20	8
x5	Avg temp(24 hours)	17	25	21	10	22	15
x6	avg temp(day)	16	25	21	10	22	17
x7	avg temp(night)	16	27	21	10	100	17
x8	avg humidity(day)	31	27	21	12	100	17
x9	avg humidity(night)	31	27	21	10	100	17
x10	avg max humidity	17	27	21	12	22	17
x11	avg mini humidity	17	27	21	10	22	17
x12	avg CO2 (ppm)	32	27	21	12	100	21
x13	Water (gift-driper)	17	20	21	12	22	21
x14	Water (gift-no)	17	20	21	12	22	21
x15	Water (gift)	17	20	21	12	22	21
x16	Water (gift)/	17	20	21	12	22	21
x17	Water (drain)/slab	17	20	21	12	22	21
x18	Water (drain)/	17	20	21	12	22	21
x19	Water (cc/J)/	17	20	21	12	22	21
x20	Water uptake/	17	20	21	12	22	21
x21	Water drain(cc/J)/	17	20	21	12	22	21
x22	Water (drain/gift)	17	20	21	12	22	21
x23	Gift EC	18	22	21	13	22	21
x24	Gift pH	18	22	21	12	22	21
x25	Slab EC	20	22	28	13	20	23
x26	Slab pH	34	25	100	12	26	27
x27	growth length	30	25	49	17	22	29
x28	cumulative growth	30	25	49	19	20	29
x29	length of leaf	30	25	49	19	20	29
x30	width of leaf	30	25	49	17	41	29
x31	no of leaves	27	25	49	15	41	21
x32	Thickness of stem	30	25	49	17	41	29
x33	height of flower	30	25	47	21	41	29
x34	no of leaves	28	25	49	21	41	21
x35	Thickness of stem	28	25	49	19	41	23
x36	height of flower	54	33	100	23	100	37
x37	Days of yield	54	33	100	23	100	37
x38	no of fruits	28	25	49	17	41	23
x39	fruit factor per unit	28	25	49	15	41	23
x40	Flowering speed	30	25	51	17	41	27
x41	set speed	31	25	53	19	41	29
x42	LAI	30	27	49	17	41	29
x43	Factor of fruit	30	25	49	25	44	25
x44	Trans1	14	25	21	12	22	8
x45	Trans2	23	27	49	17	41	8
x46	PED	28	12	49	17	41	29
x47	avg weight/unit	15	22	45	27	44	25
Overall Percentage		24	23	37	14	39	22

TABLE 5. Data Type,mean stranded deviation and Need of more clarification

Attribute Name	Attribute id	Data type	mean	standard deviation
Yield	y	Categorical	1.42	1.19
Av.outside temp.24h	x1	Continues	10.98	8.71
High temp	x2	Continues	19.97	10.30
Low temp	x3	Continues	3.44	8.81
24h Radiation sum(J)	x4	Continues	9066.50	4650.04
Av temp.24h	x5	Continues	16.85	6.99
Av temp.(day)	x6	Continues	21.15	9.06
Av temp.(night)	x7	Continues	12.32	7.78
Av hum.(day)	x8	Continues	53.92	34.56
Av hum.(night)	x9	Continues	62.47	38.89
Max hum	x10	Continues	83.85	33.18
Min hum	x11	Continues	44.70	21.01
CO2 level day(ppm)	x12	Continues	299.07	210.26
Water (gift-dripper)	x13	Continues	70.55	40.05
Water (gift-no)	x14	Continues	90.53	63.97
Water (gift)	x15	Continues	6549.72	3943.74
Water (gift)/	x16	Continues	16991.90	10246.02
Water (drain)/	x17	Continues	12086.09	9477.58
Water (drain)/	x18	Continues	6087.01	4761.83
Water (cc/J)/	x19	Continues	1.74	0.81
Water uptake/	x20	Continues	1.15	0.56
Water drain(cc/J)/	x21	Continues	0.59	0.36
Water (drain/gift)	x22	Continues	0.28	0.16
Gift EC	x23	Continues	2.15	0.94
Gift pH	x24	Continues	4.90	2.04
SlabEC	x25	Continues	3.43	1.71
SlabpH	x26	Continues	6.62	42.11
length of growth	x27	Continues	14.76	13.16
cumu. len. of growth	x28	Continues	285.91	246.58
length leaf	x29	Continues	27.17	17.87
width of leaf	x30	Categorical	23.75	16.77
no of leaves	x31	Categorical	12.69	7.65
thickness of stem	x32	Categorical	6.96	4.47
height of flower	x33	Categorical	15.95	11.33
blooming gr	x34		12.34	9.71
begin gr	x35		11.74	9.44
yield gr	x36		5.87	7.70
days of yield	x37	Categorical	14.67	16.80
no of fruits	x38	Categorical	12.50	8.35
no of fruits	x39	Categorical	32.28	21.05
Flowering speed	x40	Continues	0.64	1.00
Set speed	x41	Continues	0.60	0.79
LAI	x42	Categorical	1.83	1.46
Factor of fruit	x43	Categorical	0.55	0.45
trans1	x44		6385.51	3407.46
trans2	x45		3848.69	2803.44
PED	x46		3836.25	2978.61
average weight/unit	x47	Categorical	143.96	88.35