Taste And Nutrients Comparison Between Local And Foreign Orange Using Data Mining

By
Mehedi Hasan
Id: 171-15-9296

G.M. Shariar Al Rumman
Id: 171-15-9277

Tania Akter Lima
Id: 171-15-9269

The Report is Presented in the Requirements of Partial Fulfillment for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD SAZZADUR AHAMED
Senior lecturer Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2021

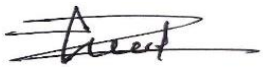
APPROVAL

This Project titled "**Taste And Nutrients Comparison Between Local And Foreign Orange Using Data Mining**", submitted by **Mehedi Hasan, ID No: 171-15-9296** , **G.M. Shariar Al Rumman, ID No: 171-15-9277** and **Tania Akter Lima, ID No: 171-15-9269** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 28th January, 2021.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Jueal Mia — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Supervised By MD SAZZADUR AHAMED Senior, lecturer Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

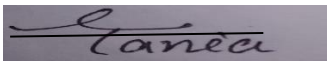
Supervised by:

MD SAZZADUR AHAMED
Senior lecturer
Department of CSE
Daffodil International University

Submitted by:

Mehedi Hasan
ID: -171-15-9296
Department of CSE
Daffodil International University

G.M Shariar Al Rumman
ID: -171-15-9277
Department of CSE
Daffodil International University



Tania Akter Lima
ID: -171-15-9269
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First of all, we express our heartfelt gratitude to Allah Almighty, for His blessing and enabling all of us the ability to successfully complete the final year without major problems.

We are extremely grateful and express our indebtedness in earnest feeling to MD SAZZADUR AHAMED, Senior Lecturer, Department of CSE Daffodil International University, Dhaka. The deep knowledge and wisdom of him in the field of "*Data Mining*" to carry out this project. His consistent patience, continual encouragement, constant and passionate supervision, constructive & helpful criticisms, valuable advices, guiding and correcting us through every stage, have made it possible to complete this project.

We wish to express our sincere gratitude to Prof. Dr. Touhid Bhuiyan, The Head of the Department of CSE, for his kind help to finish our project. And, also to other faculty member and the staff of CSE department of Daffodil International University.

Our heartfelt gratitude for our entire course mates in Daffodil International University, who took part in helping us in various ways to complete the course work.

Last but not theleast, we express our acknowledgement,with due respect, to the continuous support of our parents, the patients they showed. And, special thanks all the people that are doing their best to keep the world working in this global pandamic.

ABSTRACT

In this paper we are talking about the difference between oranges produced locally and in foreign lands. We are comparing nutrient components which tells us about the taste of the orange. As we are producing orange locally and its prospect in agriculture has bright future. But the first obstacle is the taste. We are prone to judge a fruit with how it tastes. Orange is the same. So, in order to find out the quality of the taste, we have collected the data about nutrition values of orange of different regions as well as locally produced orange. The accuracy of the collected data is more than 70% and deemed viable for the study. The lack of previous similar works and the current pandemic situation had impeded us in collecting more sophisticated data. But we have been able to collect sufficient datasets. Now dividing the data in two distinct tabular datasets, each containing around 1000 datasets, we compare the values using data mining to distinguish the differences which ultimately let us deduce the apparent taste of the orange produced in local and foreign countries.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Orange is from the family Rutaceae of citrus species fruit. The most known and famous orange is sweet orange (Citrus × sinensis). It is produced and consumed worldwide. For nutrition and especially for daily vitamin source, orange is an ideal fruit. It is the hybrid between pomelo and mandarin, two of the few ancient fruit species that survived the ice age. Sweet orange had its genome fully sequenced. Orange originated in the south and south-eastern regions of Asia consisting of the Southern regions of China, Northeast India, and Myanmar. Of course, it encompasses Bangladesh. The earliest mention of orange dates back to Chinese literature of the third century B.C. By the late twentieth century, the orange fruit tree became the most cultivated fruit tree whole worldwide. In Bangladesh, orange was cultivated since ancient times. But as time passed the quality of fruit has diminished, especially in the recent two decades we can feel the gradual decline of the taste of local orange from other foreign cultivated oranges. But the exact differences between the two are not recorded. Hence, we collected the nutrient components value of local and foreign oranges. Using data mining we want to compare the values and find out the differences we have there. By doing this we will know where to improve which will help orange production of local to increase in quality.

1.2 Motivation

From an early age, we have consumed orange as it is considerably easier to afford. But as time passed, we also noticed the obvious decline in taste of the orange. But that was locally produced. The sweetness and sourness were not what it used to be. This implied that the quality of the production was decreasing.

Here, we found that the quality of the most foreign cultivated oranges had better quality and taste. This aspired us to find out the differences in them.

Research of orange in the field of computer science is not that progressive, despite having vast research of orange in agricultural studies, food and nutrition studies, and other such. So, we decided to use orange as our research material. And then we wanted to find out the cause of difference of taste between oranges of local and abroad using data mining and analyzing.

1.3 Objective

Our research topic will be to find out the fundamental differences of orange that was cultivated locally and in foreign countries. Then we want to distinguish the facts that make local oranges taste different from foreign ones.The knowledge gained will be useful for future production quality increment and to achieve a quality of taste that is globally recognized. The farmers and orchard owners in our countries will be able to understand the quality differences between oranges. This will encourage them to produce better oranges. Have the orange growth increase nationally to contribute to the economy.

1.4 Expected outcome

Pointing the differences of oranges that were produced locally and in abroad. Learn the aspects that need to be improved which will contribute to understand about orange produced locally. Finding an optimal standard for global quality for orange that we can produce.

1.5 Research query

To what degree we can find differences of the oranges and how much we can improve the nutrient values?

1.6 Report Layout

The report is arranged as per below:

First Chapter discussed about introduction of project, motivation, research Questions, and expected outcome.

Second Chapter discussed about literature review and Challenge.

Third Chapter include Research Methodology.

Fourth Chapter includes Discussing Results and Analysis.

Fifth Chapter includes Summary and Conclusion.

CHAPTER 2

Literature Review

2.1 Introduction

We will be taking about previous researches done by other scientist that are similar to this, and about related works. The works of data analysis of orange component is very rare. So, there is a visible lack of works but orange related works are in numerus that passively helped us in our study.

In Challenge part, we will discuss about the limitations we faced primarily. Then the acceptance of data and its ingenuity.

2.2 Reviewing Literature

Juan I. Valiente, L. Gene Albrigo et.al, [1]. discussed about low temperature, crop load, and bud age on flower bud induction and orange production, the observation was done from 1999 to 2000 on orange during winter to study on naturally occurring flower bud induction. They studied on [Citrus sinensis (L.) Osbeck] or commonly known as sweet orange to determine flower induction in the seasonal climate of winter. They observed the condition of winter season, crop load, and bud age for flower bud induction. Their study was on the complex process of citrus flowering. In the winery temperature of 11° to 15° the floral intensity was increased. But the present of crop load present reduced flowering to average of 41.5%.

Maribela Pestana, Pedro Beja, Pedro José Correia, Amarilis De Varennes and Eugénio Araújo Faria et.al. [3]. studied with a field experiment to determine if flower nutrient compositions can be used to indicate fruit quality. The study was conducted in an orange orchard on a calcareous soil in southern Portugal during the three seasons from 1996 to 1999. From 30 trees the composition of flower was measured and after harvesting the quality of fruits were recorded. Principle component analysis was used for patterns of covariation in flower nutrient concentrations and the fruit quality variables were evaluated. Then regression model was developed for fruit quality variables to flower nutrient composition with stepwise selection procedures. With the predictive power of the regression model, they were able to determine that the nutrient composition of flowers at full bloom could be used to predict the fruit quality variables, fresh fruit mass, and maturation index in coming years while establishing reference values for the nutrient composition of flowers based on the measurements that was made in trees that produced large fruit.

T. Turner and B. Burri et.al. [4]. They, in their paper chose to study the nutritional benefits of citrus fruit consumption. The components were described in their works along with its health benefits which also includes oranges. They collected data from U.S. Department of Agriculture (USDA) and Food and Agriculture Organization (FAO). From the sources they were able to collect respective regional data and make analysis of the components for its benefits on health

2.3 Challenge

Like all data analysis researches, the main challenge and problem we faced was the data collection. As there was considerably lack of such research where the specific data of orange components were recorded both nationally and foreign. But in order to compare we needed a considerably large amount of data.

However, the most unexpected and perhaps the biggest problem we faced was the global pandemic of COVID-19. Because of this unprecedented situation the world came into temporary halt. We were affected in this too. Data collection almost became impossible. The primary intention of on-site experimentation and data collection was not viable because of various lockdown and closed borders. Also, databases with complete and extended period of component value data of oranges were hard to get. To device datasets we first collected the scattered data and patched it. Then we estimated the data to improve the datasets for an estimated value considering that the accuracy must not get lower than 70%.

CHAPTER 3

Methodology

3.1 Introduction

Here in methodology portion, the thing we will be discussing are about data collection method, collection procedure, preparation, algorithms, statistical analysis and implementation requirements. That is the main part of our report. Firstly, we are discussing how we collected or/and managed our data from various sources. We are also discussing about pre-procedure, proposed methodology and flow chart of our project. Then we will be done with brief but clear concept of our project and research.

3.2 Major Nutrients of Orange

Carbohydrates: The central nutrient of any organic factor. Made from carbon, hydrogen and oxygen. The primary source of organic energy. Orange has sugar and dietary fiber as carbohydrates in components. Of course, sugar is responsible for sweetness.

Protein: It is the essential nutrient component of animal physiology. Human body requires protein the most to function naturally. It is the composition of amino acids in polymer structure.

Fat: The fatty acids. Orange contains very little part of fatty acid but they are healthy for body.

Vitamins: The most crucial part of any edible fruit is its vitamin values. Vitamins are required for normal functioning of organic metabolism. Vitamin components or vitamers are present in orange. The most notable one is Vitamin-C (ascorbic acid, responsible the taste of sourness). Other vitamins are vitamin A, B-complex, E.

Minerals: The nutrient component that cannot be created biochemically in the body is known as mineral. Humans have to consume it from other outer sources such as salts. It is also essential for metabolic functioning. There are quite a few essential minerals present in orange such as, iron, calcium, Magnesium, Manganese, Potassium, and others.

Other: The extra component is the water. The orange pulp is consisting of mostly water where other components are dissolved in water.

3.3 Data Collection Procedure

This phase was the most cumbersome phase. The challenges we had to face was unprecedented due to the global pandemic situation. The primary methods of physically collecting data were unavailable. Only online data collection was viable. But data was scattered in various formats and in various types.

We collected all the scattered data that was possible for processing and turned in better and complete datasets. In this method we were able to make the of datasets that consists of the foreign orange data.

For the Bangladeshi orange we again faced the same problem. The nationwide lockdowns and extended periods of closedown of different facilities hindered on site data collection. Primary data was received from agriculture office, Khamarbari and Agricultural Universities of Bangladesh. From that we understood the layout of the data. Next, we again collected local orange data and then completed the dataset.

3.4 Data Processing:

At the data processing phase data processing we were concerned about how many types of data we got and how many formats we have gotten.

We got ourselves several types of data. We had some raw data, some were online based website data, then some were pdf file data and quite a few other formats of data that were scattered. To process this, we chose to combine and record the data in a common and generalized format which is in “xlsx” format while running the algorithm in “csv” format. Then with cleaning, tuning and other phases of processing, we got 1000 data in both of the two tabular datasets.

Data Format Sample - Local

	A	B	C	D	E	F	G	H	I
1	Energy	Sugars	Dietary fibre	Fat	Protein	Vitamin A	Beta-Carotene	Thiamine(B1)	Riboflavin(B2)
2	225 kj	10.58 g	1.9 g	0.37 g	0.80 g	30 µg	155 µg	0.060 mg	0.030 mg
3	223 kj	10.59 g	1.7 g	0.29 g	0.91 g	34 µg	145 µg	0.051 mg	0.039 mg
4	220 kj	11.02 g	1.5 g	0.34 g	0.83 g	33 µg	165 µg	0.058 mg	0.037 mg
5	219 kj	10.55 g	2.0 g	0.29 g	0.87 g	37 µg	155 µg	0.050 mg	0.033 mg
6	224 kj	10.52 g	2.3 g	0.32 g	0.84 g	39 µg	151 µg	0.057 mg	0.031 mg
7	222 kj	11.01 g	1.6 g	0.31 g	0.89 g	31 µg	158 µg	0.063 mg	0.034 mg
8	220 kj	10.50 g	1.0 g	0.30 g	0.81 g	28 µg	150 µg	0.059 mg	0.035 mg
9	225 kj	10.57 g	1.2 g	0.39 g	0.78 g	30 µg	145 µg	0.057 mg	0.032 mg
10	227 kj	11.08 g	1.4 g	0.37 g	0.92 g	35 µg	152 µg	0.055 mg	0.030 mg
11	218 kj	10.52 g	1.7 g	0.35 g	0.87 g	33 µg	157 µg	0.050 mg	0.031 mg
12	221 kj	10.50 g	1.3 g	0.30 g	0.84 g	29 µg	154 µg	0.054 mg	0.039 mg
13	229 kj	10.55 g	1.8 g	0.32 g	0.86 g	40 µg	159 µg	0.058 mg	0.037 mg
14	230 kj	11.00 g	1.0 g	0.33 g	0.89 g	37 µg	145 µg	0.050mg	0.033 mg
15	217 kj	10.58 g	2.2 g	0.28 g	0.80 g	33 µg	150 µg	0.059 mg	0.038 mg
16	222 kj	10.56 g	1.5 g	0.40 g	0.82 g	30 µg	149 µg	0.054 mg	0.032 mg
17	230 kj	10.59 g	1.3 g	0.38 g	0.80 g	32 µg	153 µg	0.051 mg	0.030 mg
18	223 kj	10.52 g	1.8 g	0.35 g	0.85 g	36 µg	158 µg	0.056 mg	0.040 mg
19	224 kj	10.48 g	2.1 g	0.37 g	0.82 g	40 µg	151 µg	0.059 mg	0.029 mg
20	228 kj	10.58 g	2.3 g	0.36 g	0.89 g	29 µg	160 µg	0.063 mg	0.035 mg
21	218 kj	10.40 g	1.8 g	0.35 g	0.80 g	34 µg	162 µg	0.060 mg	0.034 mg
22	236 kj	10.47 g	1.2 g	0.31 g	0.78 g	37 µg	154 µg	0.064 mg	0.032 mg
23	224 kj	10.49 g	1.7 g	0.29 g	0.87 g	32 µg	157 µg	0.059 mg	0.037 mg
24	223 kj	11.08 g	1.4 g	0.40 g	0.89 g	34 µg	156 µg	0.055 mg	0.039 mg
25	231 kj	11.00 g	1.9 g	0.37 g	0.80 g	30 µg	155 µg	0.058 mg	0.031 mg
26	225 kj	10.52 g	1.2 g	0.32 g	0.85 g	33 µg	153 µg	0.051 mg	0.033 mg
27	227 kj	10.48 g	1.0 g	0.31 g	0.82 g	36 µg	148 µg	0.060 mg	0.034 mg
28	230 kj	11.08 g	1.5 g	0.30 g	0.87 g	44 µg	150 µg	0.057 mg	0.036 mg
29	235 kj	10.59 g	1.7 g	0.36 g	0.80 g	42 µg	161 µg	0.053 mg	0.035 mg
30	239 kj	10.38 g	1.1 g	0.38 g	0.83 g	37 µg	158 µg	0.052 mg	0.038 mg
31	220 kj	10.58 g	1.3 g	0.30 g	0.82 g	34 µg	152 µg	0.055 mg	0.032 mg
32	219 kj	10.50 g	1.8 g	0.31 g	0.84 g	33 µg	154 µg	0.057 mg	0.030 mg
33	223 kj	10.55 g	1.4 g	0.42 g	0.87 g	30 µg	156 µg	0.059 mg	0.040 mg
34	226 kj	10.48 g	1.0 g	0.27 g	0.80 g	44 µg	157 µg	0.051 mg	0.029 mg
35	229 kj	10.59 g	1.8 g	0.37 g	0.83 g	29 µg	153 µg	0.056 mg	0.037 mg
36	232 kj	10.46 g	1.2 g	0.35 g	0.88 g	27 µg	155 µg	0.057 mg	0.033 mg
37	219 kj	10.48 g	2.0 g	0.34 g	0.90 g	34 µg	158 µg	0.053 mg	0.031 mg
38	223 kj	10.58 g	1.4 g	0.36 g	0.81 g	30 µg	150 µg	0.055 mg	0.039 mg

Figure 3.4.1: Sample-1

Data Format Sample – Foreign

	A	B	C	D	E	F	G	H	I	J	K
1	Country	Energy	Sugars	Dietary	Fat	Protein	Beta-Caro	Vitamin	Thiamine(B1)	Riboflavin(B2	Niacin(B3)
2	USA	190 kj	9.35 g	2.4 g	0.19 g	1.00 g	155 µg	11 µg	0.097 mg	0.04 mg	0.292 mg
3	Brazil	223 kj	9.07 g	2.10 g	0.41 g	0.73 g	150µg	16 µg	0.101 mg	0.10 mg	0.401 mg
4	USA	191 kj	9.45 g	2.9 g	0.13 g	0.97 g	153µg	14 µg	0.097 mg	0.04 mg	0.294 mg
5	USA	195 kj	9.39 g	2.0 g	0.11 g	0.98 g	149µg	16 µg	0.085 mg	0.04 mg	0.287 mg
6	China	205 kj	10.09 g	2.50 g	0.40 g	0.99 g	157µg	19 µg	0.097 mg	0.050 mg	0.279 mg
7	USA	190 kj	9.35 g	2.6 g	0.21 g	0.96 g	153µg	17 µg	0.086 mg	0.04 mg	0.282 mg
8	India	197 kj	9.15 g	2.4 g	0.29 g	0.79 g	155 µg	20 µg	0.102 mg	0.09 mg	0.409 mg
9	USA	197 kj	9.38 g	2.1 g	0.11 g	1.04 g	150µg	19 µg	0.085 mg	0.04 mg	0.286 mg
10	USA	189 kj	9.30 g	3.1 g	0.12 g	1.09 g	153µg	19 µg	0.089 mg	0.04 mg	0.289 mg
11	USA	193 kj	9.45 g	2.4 g	0.32 g	0.94 g	149µg	10 µg	0.090 mg	0.04 mg	0.282 mg
12	China	203 kj	10.10 g	2.52 g	0.38 g	0.97 g	157µg	18 µg	0.086 mg	0.047 mg	0.273 mg
13	USA	199 kj	9.34 g	3.0 g	0.14 g	0.97 g	153µg	21 µg	0.086 mg	0.04 mg	0.287 mg
14	USA	187 kj	9.31 g	2.4 g	0.10 g	0.95 g	155 µg	19 µg	0.084 mg	0.04 mg	0.285 mg
15	Brazil	223 kj	9.15 g	2.15 g	0.31 g	0.86 g	150µg	15 µg	0.107 mg	0.10 mg	0.405 mg
16	USA	197 kj	9.37 g	2.1 g	0.19 g	0.94 g	153µg	15 µg	0.097 mg	0.04 mg	0.292 mg
17	China	202 kj	10.08 g	2.53 g	0.30 g	1.04 g	149µg	18 µg	0.089 mg	0.045 mg	0.276 mg
18	India	192 kj	9.11 g	2.0 g	0.27 g	0.79 g	157µg	16 µg	0.105 mg	0.06 mg	0.401 mg
19	USA	188 kj	9.45 g	2.9 g	0.09 g	0.99 g	153µg	18 µg	0.095 mg	0.04 mg	0.288 mg
20	India	190 kj	9.19 g	2.1 g	0.28 g	0.72 g	150µg	15 µg	0.109 mg	0.03 mg	0.407 mg
21	China	201 kj	10.09 g	2.51 g	0.37 g	0.99 g	153µg	19 µg	0.080 mg	0.049 mg	0.274 mg
22	USA	193 kj	9.36 g	3.0 g	0.22 g	0.90 g	149µg	17 µg	0.087 mg	0.04 mg	0.290 mg
23	USA	199 kj	9.34 g	2.1 g	0.12 g	1.09 g	157µg	11 µg	0.088 mg	0.04 mg	0.287 mg
24	Brazil	223 kj	9.07 g	2.10 g	0.43 g	0.73 g	153µg	15 µg	0.103 mg	0.11 mg	0.406 mg
25	USA	195 kj	9.30 g	2.6 g	0.10 g	0.99 g	155 µg	19 µg	0.084 mg	0.04 mg	0.287 mg
26	USA	196 kj	9.36 g	2.0 g	0.13 g	0.94 g	150µg	21 µg	0.082 mg	0.04 mg	0.283 mg
27	USA	191 kj	9.39 g	2.5 g	0.15 g	1.05 g	153µg	23 µg	0.089 mg	0.04 mg	0.289 mg
28	Brazil	227 kj	9.09 g	2.11 g	0.51 g	0.79 g	149µg	19 µg	0.108 mg	0.11 mg	0.409 mg
29	India	195 kj	9.18 g	3.4 g	0.28 g	0.76 g	157µg	19 µg	0.107 mg	0.09 mg	0.409 mg
30	USA	190 kj	9.38 g	2.9 g	0.18g	1.00 g	153µg	09 µg	0.081 mg	0.04 mg	0.292 mg
31	USA	197 kj	9.31 g	2.3 g	0.22 g	0.94 g	155 µg	18 µg	0.098 mg	0.04 mg	0.282 mg
32	Brazil	224 kj	9.15 g	2.19 g	0.39g	0.79 g	150µg	19 µg	0.109 mg	0.14 mg	0.404 mg
33	USA	198 kj	9.30 g	2.0 g	0.12 g	0.93 g	153µg	13 µg	0.090 mg	0.04 mg	0.293 mg
34	China	201 kj	10.03 g	2.45 g	0.30 g	1.06 g	149µg	18 µg	0.085 mg	0.047 mg	0.274 mg
35	USA	193 kj	9.40 g	2.9 g	0.20 g	0.92 g	157µg	22 µg	0.099mg	0.04 mg	0.289 mg
36	USA	190 kj	9.41 g	2.0 g	0.11 g	0.97 g	153µg	19 µg	0.084 mg	0.04 mg	0.282 mg
37	Brazil	222 kj	9.09 g	2.10 g	0.43 g	0.70 g	155 µg	15 µg	0.104 mg	0.12 mg	0.402 mg

Figure 3.4.2: Sample-2

Here are the sample of data sets containing the nutrient values of local and foreign oranges.

3.5 Data Pre-Processing:

After collecting all the data, we processed it for our implementation. We have removed or generalized/normalized all null data and trained it as our model to find the maximum accuracy. As our algorithms are using KNN classifier primarily, majority of the data trained are used for testing. So, most of the data trained is used for testing to get the expected outcomes.

3.6 K Nearest Neighbor (KNN)

KNN is the most basic and simplest separation algorithm. Also, it is one of the most widely used learning algorithms. It is the a non-parametric, instance-based, or more commonly, the algorithm for lazy learning. KNN aims to use a database where data points are divided into several categories and used to predict the division of a new sample point. Here, non-parametric means that KNN does not make any assumptions in terms of basic data distribution. As a result, we can create a model from the data. Its usefulness depends on the fact that in the real world, most details do not follow common theories.

For this reason, the KNN algorithm is our first choice for classification research where there is little or no prior knowledge about distribution data. Also, by being a "Lazy Algorithm", KNN does not use standard data points that reduce the training data level or reduce it. KNN keeps all the details of the test phase training.

Now, KNN acts as an object classified based on the closest number of the object types. The item is given the most common category among the nearest neighbors k (k to be the integer).

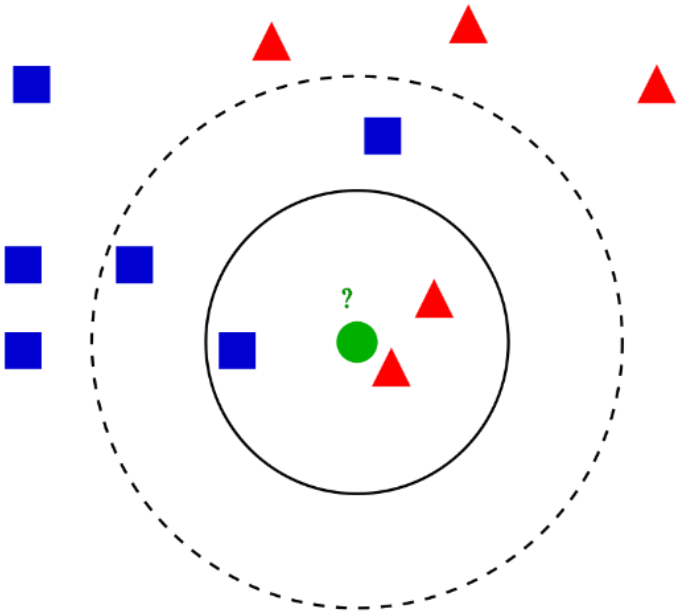


Fig:3.6.1 K-NN classification Example

As displayed in the figure above, the green dot (test sample) should be placed in the red triangular section or in the blue squares. If k = 3 (solid circle), it is given a red triangle because there are only 2 triangles and only 1 square inside the inner circle. If k = 5 (dash circle), it is given blue squares as 3 squares compared to 2 triangles inside the outer circle where there are many squares.

3.7 Naive Bayes Algorithm

Our next choice of algorithm is the Naive Bayes algorithm.Based on Bayes' Theorem.

The Naive Bayes algorithm is a classification method that works by assuming that the scale elements are independent of each other. Assuming the presence of a particular feature or forecast in the class is not related to the presence of any other features or predictions.

Bayes theorem calculates the posterior probability P (c | x) from P (c); P (x); and P (x | c). See figure below–

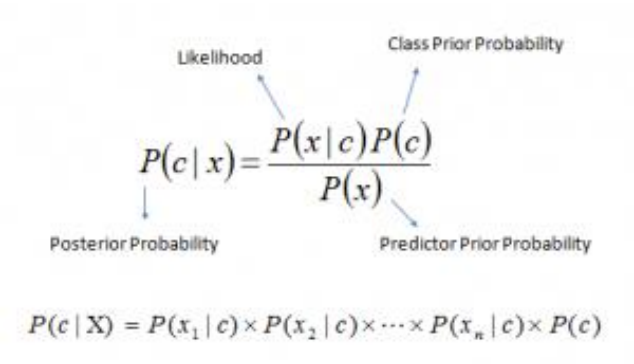


Fig :3.7.1 Naive Bayesian Equation

Here, in the equation

P(c|x) = posterior probability of class (c,target) given predictor (x, attributes).

P (c) = probability prior to phase.

P (x | c) = probability of a given predictive phase given.

P (x) = pre-forecast probability.

The algorithm works easily. In three steps it can be explained.

Step one: Change the data set in the frequency table.

Step Two: By finding opportunities, Create an Opportunity Table.

Step Three: You should use the "Naive Bayesian Equation" to calculate the background opportunities for each class. The category with the highest probability in the background should be the result of the forecast.

3.8 Proposed Methodology

Methodology –

We can analyze comparison using classification techniques. Here, we are using KNN and Naïve Bayes classification algorithms. As the use of KNN in data mining is very basic and we can have use it while the training phase can be overlooked. As a lazy algorithm KNN uses most if not all of the training data for test data. So, using KNN is easier in this case. Naïve Bayes Classifier assumes independent predictor and features of a measurement. It usage the Naïve Bayesian equalization to compute the probability of succession.

Algorithms –

With the **KNN** algorithm- to avoid over fitting, we split our data into test and test phases, which gives us a better idea of how our algorithm performed during testing. In this way, our algorithm is tested on invisible data.

We will fit/train the separator on the training set and make predictions on the test set. We will then compare the prediction with a well-known label. Scikit-learn provides a data center on the train and test system using the train_test_split method. Also, we create a trial set the size of about 20% of the data
To perform classification and testing, we will use the following code:

```
In [6]: #importing train_test_split
        from sklearn.model_selection import train_test_split

In [7]: X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=45, stratify=y)
```

We also calculated the matrix for the confusion. The confusion matrix is a commonly used table to describe the performance of a subdivision in a test data set where true values are known. Scikit-learn provides a place to calculate the confusion matrix using the confusion_matrix method. We using Scikit-Learn to calculate the confusion matrix.

we will use the following code for Local data-

```
In [13]: #import confusion_matrix
         from sklearn.metrics import confusion_matrix

In [14]: #let us get the predictions using the classifier we had fit above
         y_pred = knn.predict(X_test)

In [15]: confusion_matrix(y_test,y_pred)

Out[15]: array([[46, 43],
                [32, 79]], dtype=int64)
```

the following code for Foreign type of data-

```
In [14]: #import confusion_matrix
         from sklearn.metrics import confusion_matrix

In [15]: #let us get the predictions using the classifier we had fit above
         y_pred = knn.predict(X_test)

In [16]: confusion_matrix(y_test,y_pred)

Out[16]: array([[37, 52],
                [34, 77]], dtype=int64)
```

Confusion matrix table for Local data:

```
In [23]: pd.crosstab(y_test, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)
```

Out[23]:

True	0	1	All
0	37	52	89
1	34	77	111
All	71	129	200

Confusion matrix table for Foreign data:

```
In [16]: pd.crosstab(y_test, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)
```

Out[16]:

True	0	1	All
0	46	43	89
1	32	79	111
All	78	122	200

Now we will compute the Classification report. It is a summary of the text with precision, recall, F1 score for each class. Actually, Scikit-learn provides a place to calculate the classification report using the classification_report method. We will calculate both local and Foreign types of data.

the following code for Local data-

```
In [24]: #import classification_report
         from sklearn.metrics import classification_report
```

```
In [25]: print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.52      0.42      0.46        89
           1       0.60      0.69      0.64       111

    accuracy                           0.57       200
   macro avg       0.56      0.55      0.55       200
weighted avg       0.56      0.57      0.56       200
```

the following code for Foreign data-

```
In [17]: #import classification_report
         from sklearn.metrics import classification_report
```

```
In [18]: print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.59      0.52      0.55        89
           1       0.65      0.71      0.68       111

    accuracy                           0.62       200
   macro avg       0.62      0.61      0.61       200
weighted avg       0.62      0.62      0.62       200
```

With the **Naïve Bayes** algorithm- On the other hand, we execute also Naïve Bayes algorithm for our local and foreign data. We split the dataset into two positions the training set and the test set. The following separates the data from the training and test set at 80:20. Here is algorithm which we execute-

```
In [79]: # Splitting the dataset into the Training set and Test set
         from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.20, random_state=1)
```

After the pre-processing step, we fitted the Naïve Bayes model to the Training set. In the above code, we have used the GaussianNB classifier to fit it to the training dataset. Below is the code for it-

```
In [84]: # Fitting Naive Bayes to the Training set
         from sklearn.naive_bayes import GaussianNB

In [85]: clf = GaussianNB()

In [86]: clf.fit(X_train_std, y_train)
Out[86]: GaussianNB()

In [87]: GaussianNB(priors=None)
Out[87]: GaussianNB()
```

Then we predicted the test set result. For this, we created a new predictor variable y_pred, and we used the predict function to make the predictions.

```
In [88]: # Predicting the Test set results
         y_pred = clf.predict(X_test_std)
```

Working Procedures –

The diagram below shows a general and overall working of the procedures

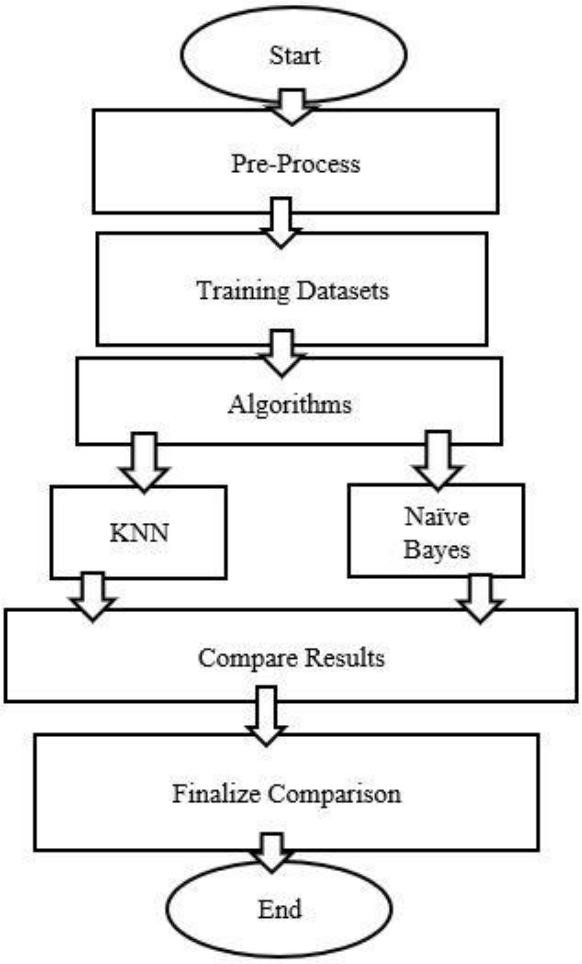


Figure 3.8.1: Working procedures diagram

Here, we are working primarily by running two different classification algorithms on the datasets. After collecting and processing data, we use KNN and Naïve Bayes algorithms on the datasets separately. There we collect the analysis of the datasets that contains the nutrient values. Then we compare and generate result based on the outcomes of the algorithms.

After gathering the data, finish the pre-processing. We get sizeable datasets to work on after processing phase. Then we run the algorithms. First, we use KNN on the datasets. We get the analysis result as below,

Final Accuracy for Local dataset-

```
In [13]: #Get accuracy. Note: In case of classification algorithms score method represents accuracy.
         knn.score(X_test,y_test)

Out[13]: 0.57
```

Final Accuracy for Foreign dataset-

```
In [12]: #Get accuracy. Note: In case of classification algorithms score method represents accuracy.
         knn.score(X_test,y_test)

Out[12]: 0.625
```

Here, it is shown the accuracy we got after the data mining on foreign and local datasets. From the analysis we can observe that the accuracy rate is around 62% for foreign dataset and 57% for local dataset by using KNN algorithm.

Now we use Naïve Bayes on the local dataset.

```
In [89]: # Making the Confusion Matrix
         from sklearn.metrics import accuracy_score

In [90]: # Final Accuracy
         accuracy_score(y_true=y_test, y_pred=y_pred)

Out[90]: 0.705
```

Local dataset shows the accuracy around 70%. Then use naïve bayes on the

foreign dataset. The result –

```
In [110]: # Making the Confusion Matrix
          from sklearn.metrics import accuracy_score

In [111]: # Final Accuracy
          accuracy_score(y_true=y_test, y_pred=y_pred)

Out[111]: 0.75
```

With the outcome that was around 75% accuracy.

Implementation Requirements:

After analyzing all the essential procedures and works, we are in need of some specific hardware and software as well as some implementing tolls for analyzing the comparison.

Required Hardware and software:

- A computer device/system with up-per configuration and/or specifications.
- The most updated operating system like widows 10 or Linux.

Tools of Implementation:

- Anaconda with Jupiter notebook.
- Python.
- Scikit-Learn.
- NumPy.
- Pandas Data Frame.
- Matplotlib.
- Microsoft Office (Excel).

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction:

After running the algorithms, we find the outcomes that shows us the comparison result. The algorithms analyze datasets. First, it shows the outcome of the analysis of the nutrient values separately. Then analyze the comparison between the datasets. Finally, we conclude based on the outcome of the analysis and comparison made.

4.2 Analyzing Result and Summarization:

Applying the algorithms, we examine the values of nutrient present in the datasets. We also use jupyter development kits and python language for testing accuracy level of our data. With these methods, we were able to generate a satisfactory level of initially more than 60% of accuracy one average.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Study Summary

Various types of agriculture play essential role in the economic sector of our country. Studies related to it always makes progress of our national development. Orange has its origin etched in this country. Also, the seasonal cycle is better for producing orange. In recent years, research on orange cultivation has risen in these regions. Researcher are using various methods to study orange cultivation prospects. But evidential works on comparing local orange to foreign orange is relatively low. So, we targeted this aspect while using data mining to solve the problems. We are using datasets with majority of trained and testing data. With this, we were able to have outcome of accuracy around 65% to 70% on average. It proves that our methodology was in the correct direction and with more substantial data we can have better outcome and result. This study will help to understand essence of orange and points to improve its taste and nutrient to achieve global standard.

5.2Conclusions

Orange cultivation is developing in our country in recent year and all types of work on orange will raise production growth and increase the quality of orange. Our study being one of the first work of such field where we try to distinguish a global standard of taste and nutrient value for locally produced orange. Our accuracy and data outcome has given enough result to progress further with the intention of betterment of national agriculture. To work better, we of course need vast amount of data which can be used to have even better and precious results.

Also finding out most optimal state for production of orange. And, ultimately contributing in developing orange related horticulture. From the study of this research, we can also work in the same way to conclude the overall quality of other fruit compared to local for finding out the lacking for a global standard.

5.3 Recommendations

Few Recommendations about this research are at first collect more component data for better accuracy and also Collect more variety of foreign data. Then Try to fill up all null values with the average trained value. Finally, try to synchronize the seasonal production to gain analysis of much higher accuracy.

5.4 Prospects for Further Study

More implementation and research prospects for future studies are adding more regional foreign data (more different countries). We want to add at least a thousand data for our future research purpose which will further increase our accuracy of data. Add new types of data concerning cultivation environment, climate, and other such essential information. We want to study to find out how we can improve the taste and quality. Would like to work with the government for researching the aspect of making locally produced oranges accepted globally.

REFERENCES

[1] Juan I. Valiente, L. Gene Albrigo. “Flower Bud Induction of Sweet Orange Trees: Effect of low temperatures, Crop Load, Bud Age.” *J. AMER. SOC. HORT. SCI. 129(2),* vol. 1, pp 158–164, March 2004.

[2] Md. Nazmul Hasan, Mohammed Raqibul Hasan, Shakhawat Hossain Foysal, Hammadul Hoque, Md. Fahim Khan, Md. Fahmid Hossain Bhuiyan, Shamsul H. Prodhan. “In-Vitro Regeneration of Citrus sinensis (L.) Osbeck from Mature Seed Derived Embryogenic Callus on Different Solid Basal Media.” *American Journal of Plant Sciences,* vol. 10, pp. 285-297, February 2019.

[3] Maribela Pestana, Pedro Beja, Pedro José Correia, Amarilis De Varennes and Eugénio Araújo Faria “Relationships between nutrient composition of flowers and fruit quality in orange trees grown in calcareous soil.” *Tree Physiology*, vol 25(6), pp 761–767, June 2005.

[4] T. Turner and B. Burri, “Potential Nutritional Benefits of Current Citrus Consumption,” *Agriculture*, vol. 3, no. 1, pp. 170–187, Mar. 2013.

[5] “*Food Surveys Research Group: Beltsville, MD*”, available at <<https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/>> last accessed on 07-01-2021 at 9:20 pm.

[6] “*FoodData Central*”, available at << https://fdc.nal.usda.gov/index.html>> last accessed on 17-07-2020 at 4:42 pm.

[7] "*Mandarin orange*", available at << https://en.wikipedia.org/wiki/Mandarin_orange>> last accessed on 23-12-2020 at 11:12 am.

[8] "*Orange (fruit)*", available at << https://en.wikipedia.org/wiki/Orange_(fruit) >> last accessed on 23-12-2020 at 11:12 am

[9] Parle Milind, Chaturvedi Dev. "Orange- range of Benefits." *International Research journal of pharmacy,* vol. 3, pp. 59-63. June 2012.

[10] "*FoodData Central*", available at << https://fdc.nal.usda.gov/fdc-app.html#/food-details/169103/nutrients>> last accessed on 08-09-2020 at 1:32 pm.

[11] "*Sweet Orange*", available at << https://www.sciencedirect.com/topics/medicine-and-dentistry/sweet-orange#:~:text=These%20fruits%20have%20different%20chemical,essential%20oils%20can%20be%20found.&text=The%20major%20component%20of%20citrus,lemon%20containing%2045%E2%80%9376%25.>> last accessed on 14-08-2020 at 8:03 pm.

[12] "*Oranges 101: Nutrition Facts and Health Benefits*", available at << https://www.healthline.com/nutrition/foods/oranges>> last accessed on 04-01-2021 at 7:11 pm.

Taste And Nutrients Comparison Between Local And Foreign Orange Using Data Mining

ORIGINALITY REPORT

15%	14%	5%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.scirp.org Internet Source	1%
4	www.edureka.co Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	export.arxiv.org Internet Source	1%
7	link.springer.com Internet Source	1%
8	K.F. Bangerth. "Floral induction in mature, perennial angiosperm fruit trees: Similarities and discrepancies with annual/biennial plants and	<1%