International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 16 (2015) pp36868-36871
© Research India Publications. http://www.ripublication.com

Data Analysis On Diet Management And Disease Prediction

Mr.T. Rajasekaran
Assistant Professor
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology,
Arasur,Coimbatore District.
rajasekaran30@gmail.com

R. Dhivakar
UG Scholar
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology,
Arasur,Coimbatore District.
divii.ragu@gmail.com

M. Gokul Anand
UG Scholar
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology,
Arasur,Coimbatore District.
gokulanand2@gmail.com

N. Arun Kumar
UG Scholar
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology,
Arasur,Coimbatore District.
arunkumar.nnn@gmail.com

Abstract-This Paper introduces a technique that will analyze the health of a person and look for the possible diseases that can occur in the near future based on the food intakes. The health can be maintained only in terms of good and balanced diet food. As people are moving towards fast foods, the intake of healthy food has been forced to a notable decline in the past few decades. The Statistical analysis can be done over the food plan of an individual and suggestions can be given to improve their diet plan. Hence, providing some provisional information and warnings about the diseases that may come to the person helps the person to get free from the diseases. The disease prediction can be done using Predictive Analysis over the food plan and their tendency to cause the diseases. It mainly aims at helping the human community to live with good immunity.

Keywords: Statistical analysis, Predictive analysis, Disease Prediction, Diet Management, J48, NLTK, Matplotlib

1. INTRODUCTION

'Health is Wealth' is a statement that is well accepted by everyone. The health organizations all around the world are in need to reduce the cost and to improve the efficiency and accuracy in disease diagnosis. In addition to this, there are several proofs that show the inefficiency and sub-optimal clinical results [1].Hence, there is a need for quick and effective mechanism for finding the diseases in the medical field. The Statistical and Predictive analysis over the data would be more helpful in bringing out accuracy and speedy disease diagnosis.

1.1 Diet Management

The changes made to the diet are the long term changes of an individual's health and so every diet must be a health-minded one [2].Even though people are educated through-out their life about the importance of diet and almost everyone is being conscious about the importance of their regular diet, the change in food pattern of the society makes their diet an unbalanced one.

The Statistical analysis over the diet of an individual analyzes the diet based on the amount of food they in-take in an average and the ingredients that are used in the food and generates a graph reporting the healthy and unhealthy factors of the diet.

1.2 Disease Prediction

Manually accounting for all the possible factors of all the disease is really a tedious task [3].Therefore, we need a simple mechanism to identify the disease and deficiencies and should ensure that it should be a speedy one. The food plan is given as the input and possible diseases are found out as the result using Predictive analysis.

1.2.1 Food Borne Diseases

The diseases that are born and spread through foods are more common and life-threatening issues all around the world. One in six Americans are hospitalized and one among forty two who were hospitalized die due to the food borne diseases [4].Hence, there is a serious relationship between the food we intake and the disease come out to us. Such diseases should be predicted before and awareness should be given to the people before consuming it without any knowledge about what it will bring in future.

2. ANALYTICS MODEL

2.1 Statistical Analytics Model

In general, Statistical analysis helps in finding out the unhidden patterns and the unrevealed correlations among the data that are difficult for the humans to interpret and find out. It is the science of collecting, exploring and presenting the large volume of data and discovering the underlying patterns and trends from it and using it in a useful way [5]. This technique can be used for diet management. The analysis is done over the diet of the user. Based on the ingredients used in the food and their chemical composition the foods consumed are classified into Healthy, Unhealthy and moderate one.

The Classification is done using the python programming with the lexical analyzing packages like NLTK (Natural Language Tool Kit) and the outputs are plotted as a graph using pylab and matplotlib packages.

2.1.1 NLTK (Natural Language Tool Kit)

The Natural Language Tool Kit is of the program modules, dataset and sometimes tutorials which support the research and teaching in computational linguistics and the processing of natural language. The NLTK is written using python and distributed under the GPL Open Source License [6].

2.1.2 Matplotlib

Matplotlib is a library used in python for 2-D graphics. The major application of matplotlib are interactive scripting and publication-quality image generation across the various user Interfaces and operating systems [7].The matplotlib is used for the generation of a graph that shows the balance of diet of the individual and a second graph which gives a comparative chart between normal and user's healthy, moderate and unhealthy food intake Levels. The graphical representation provides an easier understanding of the user diet and brings a good conscious to the user about his diet levels.

2.2 Predictive Analytics Model

Predictive analysis model uses a variety of techniques from statistics, data mining and game theory and analyze the current and past facts to make the correct prediction. The predictive model takes the input from the historical facts which are called as the training dataset and classifies the current data items. This is done with the help of J48 algorithm. The J48 algorithm is used as a statistical classifier, which generates a decision tree based on the input[10].

3. PROBLEMS TO BE SOLVED

Data collection about each and every food type and also about the ingredients that are present in the food and the quality of the ingredients also should be analyzed for making the smarter decisions.

The Consumption of certain foods alone does not lead to the diseases. The consumption of foods above the limit increases the certain components level in the body that will lead to the diseases. Hence, proper monitoring of the constituents of the food is also very important.

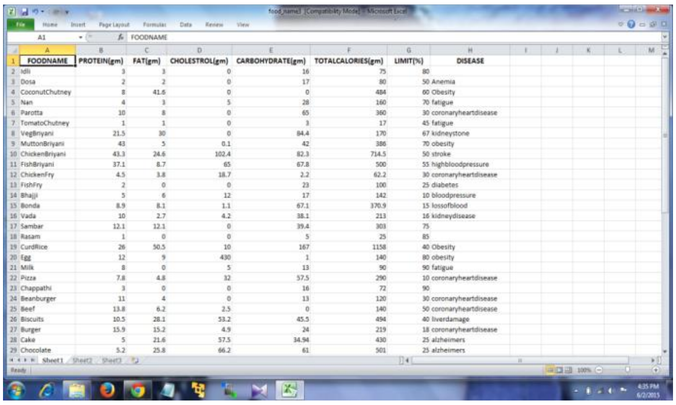
Finding the proper algorithms is also a major factor in the prediction. Hence an algorithm which is more efficient and optimal should be found out and used for optimal solution.

4. METHEDOLOGY

The importance of the prediction of diseases before the occurrence plays a vital role in increasing the life span of human. There are several mechanisms in predicting the future diseases using the art-of-possible predictions method. [9]

4.1 Data Collection

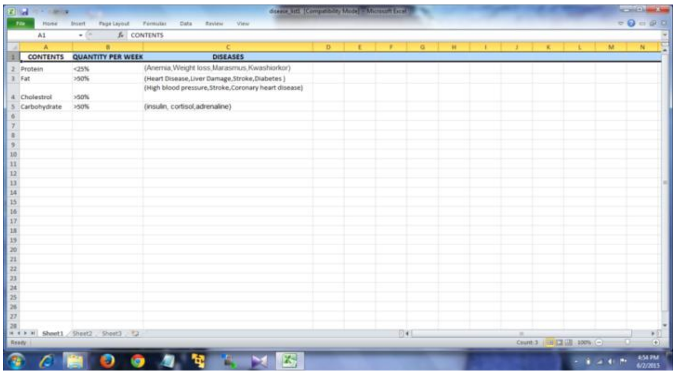
The overall process begins from the collection of datasets which consists of variety of foods and their constituents and the chemical composition. It is shown in figure 1. Another data set is needed, which includes the information like quantity of chemical components needed and their effects if it is taken in excess. It is shown in figure 2.



A1 FOODNAME

	A	B	C	D	E	F	G	H
1	FOODNAME	PROTEIN(gm)	FAT(gm)	CHOLESTROL(gm)	CARBOHYDRATE(gm)	TOTALCALORIES(gm)	LIMIT(%)	DISEASE
2	Idli	3	3	0	16	75	80	
3	Dosa	2	2	0	17	80	50	Anemia
4	CoconutChutney	8	41.6	0	0	484	60	Obesity
5	Nan	4	3	5	28	160	70	fatigue
6	Parotta	10	8	0	65	360	30	coronaryheartdisease
7	TomatoChutney	1	1	0	3	17	45	fatigue
8	VegBriyani	21.5	30	0	84.4	170	67	kidneystone
9	MuttonBriyani	43	5	0.1	42	386	70	obesity
10	ChickenBriyani	43.3	24.6	102.4	82.3	714.5	50	stroke
11	FishBriyani	37.1	8.7	65	67.8	500	55	highbloodpressure
12	ChickenFry	4.5	3.8	18.7	2.2	62.2	30	coronaryheartdisease
13	FishFry	2	0	0	23	100	25	diabetes
14	Bhajji	5	6	12	17	142	10	bloodpressure
15	Bonda	8.9	8.1	1.1	67.1	370.9	15	lossofblood
16	Vada	10	2.7	4.2	38.1	213	16	kidneydisease
17	Sambar	12.1	12.1	0	39.4	303	75	
18	Rasam	1	0	0	5	25	85	
19	CurdRice	26	50.5	10	167	1158	40	Obesity
20	Egg	12	9	430	1	140	80	obesity
21	Milk	8	0	5	13	90	90	fatigue
22	Pizza	7.8	4.8	32	57.5	290	10	coronaryheartdisease
23	Chappathi	3	0	0	16	72	90	
24	Beanburger	11	4	0	13	120	30	coronaryheartdisease
25	Beef	13.8	6.2	2.5	0	140	50	coronaryheartdisease
26	Biscuits	10.5	28.1	53.2	45.5	494	40	liverdamage
27	Burger	15.9	15.2	4.9	24	219	18	coronaryheartdisease
28	Cake	5	21.6	57.5	34.94	430	25	alzheimers
29	Chocolate	5.2	25.8	66.2	61	501	25	alzheimers

Figure 1. Dataset for foods and their constituents and the chemical composition of the constituents.



A1 CONTENTS

	A	B	C
1	CONTENTS	QUANTITY PER WEEK	DISEASES
2	Protein	<25%	(Anemia,Weight loss,Marasmus,Kwashiorkor)
3	Fat	>50%	(Heart Disease,Liver Damage,Stroke,Diabetes) (High blood pressure,Stroke,Coronary heart disease)
4	Cholestrol	>50%	
5	Carbohydrate	>50%	(insulin, cortisol,adrenaline)

Figure2. Dataset forquantity of chemical components needed and their effects

4.2 Statistical Analysis Using Python:

The Statistical analysis is performed over the diet of the user which is read from the user. The diet is analyzed using the lexical analysis packages NLTK and the classification of the foods are done as healthy, moderate and unhealthy food. From the results a graph is plotted representing the diet condition of the user that is shown in figure 3. The outputs are also shown about their healthy, moderate and unhealthy food. Finally diet of the user is shown in figure 4.

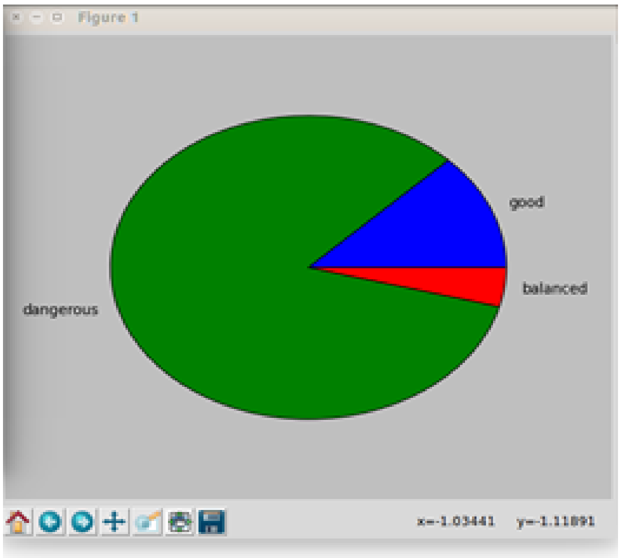


Figure 3. Represents the food intake as dangerous,balanced and good.

Based upon the classifications of the food that is obtained from the NTLK, a bar-chart is drawn using the matplotlib library, for representing a graph that shows the comparison between the amount of healthy, moderate and unhealthy food that an individual can intake and the amount that the user intakes.

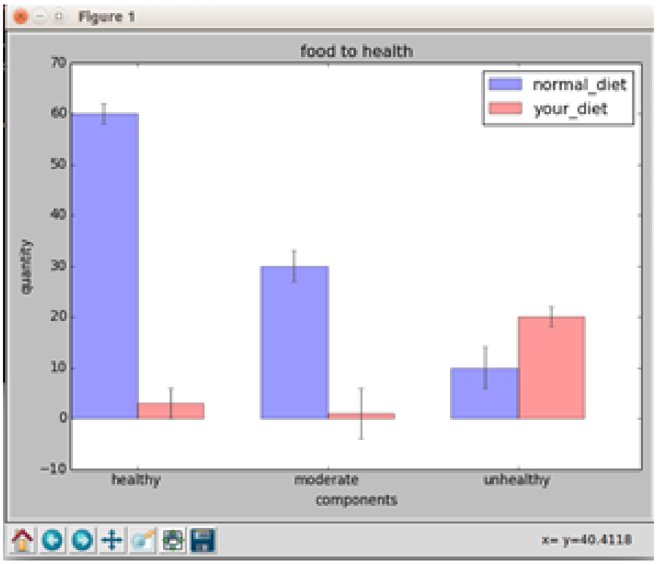


Figure 4. Represents the diet of user in healthy, moderate and unhealthy

These two charts are very helpful in checking the balanced diet of a person.

4.3 Disease Prediction Using J48 Algorithm

The decision tree is used as a classification mechanism. The classification mechanism followed by the J48 algorithm is building of class models from a training set of data that contains a class labels. Output is shown in figure 6.

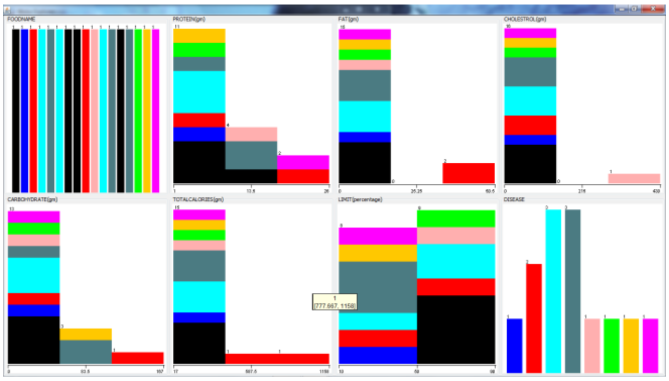
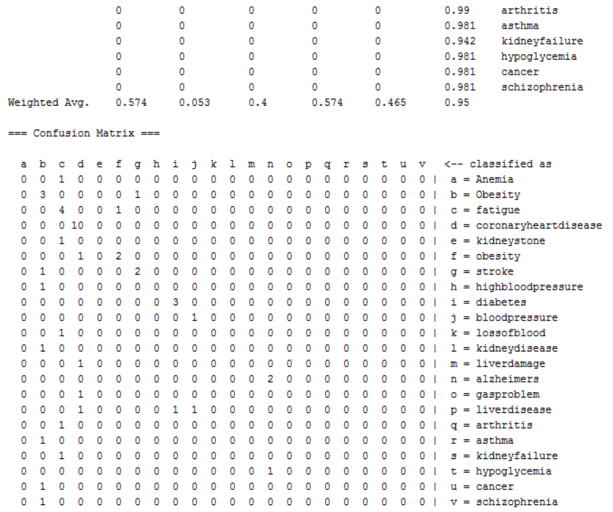


Figure5. Represents the quantity of instance in each attributes

The J48 decision tree algorithm is used to find the way in which components in the food behaves for a number of instances which are shown in figure 6 and based on the training dataset, the classes for the newly formed instances are also found.



```
                 0         0          0         0         0          0.99     arthritis
                 0         0          0         0         0          0.981    asthma
                 0         0          0         0         0          0.942    kidneyfailure
                 0         0          0         0         0          0.981    hypoglycemia
                 0         0          0         0         0          0.981    cancer
                 0         0          0         0         0          0.981    schizophrenia
Weighted Avg.    0.574     0.053      0.4       0.574     0.465      0.95

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v   <-- classified as
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = Anemia
  0  3  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = Obesity
  0  0  4  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = fatigue
  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  d = coronaryheartdisease
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = kidneystone
  0  0  0  1  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = obesity
  0  1  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  g = stroke
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  h = highbloodpressure
  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0 |  i = diabetes
  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0 |  j = bloodpressure
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  k = lossofblood
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  l = kidneydisease
  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  m = liverdamage
  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0 |  n = alzheimers
  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  o = gasproblem
  0  0  0  1  0  0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0 |  p = liverdisease
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  q = arthritis
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  r = asthma
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  s = kidneyfailure
  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0 |  t = hypoglycemia
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  u = cancer
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  v = schizophrenia
```

Figure 6. Classification using the J48 Algorithm

The predictions of the diseases are made by adding the instances generated for any food based on their constituents to a disease-named class. The graph is shown at figure 7,where the x-plot represents the food name and the y-plot represents the associated diseases. Therefore, it represents the predicted disease of an individual.

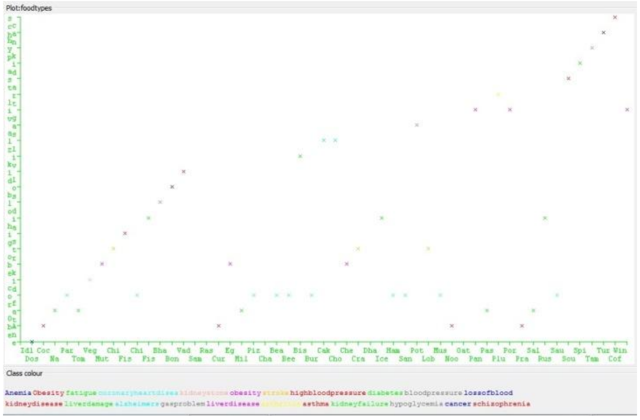


Figure 7. Represents the Food name and its associated diseases.

5. CONCLUSION

The diet management using statistical analysis helps the user to have a keen observation regarding their health. It also eliminates the need of monitoring the diet for long days. This would be very helpful for a common man to get awareness about his diet. The disease prediction helps to find out the disease that has occurred or may occur without any medical tests. Since, it is more economic and a powerful mechanism in finding out the diseases, this would be a new revolution in health sector and increase the relationship between the data and science even more closely.

6. FUTURE WORKS

Implementing the system in mobile applications. For example, in windows and in android platforms, such that every user can have the system with them wherever they go. Suggesting a certain healthy food schedule for the user and suggesting the food for a particular time to balance the diet. It can be implemented in hospitals, schools, hostels, hotels, etc., to show their food diet and secured from diseases. It will have a user login to store records of individual users, so that it helps for their future reference.

7. REFERENCES

[1] The Value of analytics in Health Care [James W. Cortada, Dan Gordon and Bill Lenihan].

[2] American Journal of Clinical Nutrition [69-373] [Lee., et al.].

[3] Predicting outbreak Severity through machine learning on disease out-break reports [Rowan Chakoumakos (Stanford University)].

[4] FOODBORNE DISEASES, Health and Research, from NAID [National Institute of Allergy and Infectious Diseases].

[5] www.sas.com, Statistical-Analysis.

[6] NLTK: Natural Language Toolkit by Steven Bird and Edward Loper (DOI:10.3115/1225403.1225421. on 21st International conference on Computational Linguistics)

[7] Matplotlib, computing in science and engineering. Volume-9, Issue: 3, pages (90-95), ISSN: 9501168; DOI: 10.1109/mCSE.2007.55

[8] www.wikipedia.com, “Predictive Analysis” 2010

[9] How to make Predictions about infectious diseases risk? Mark wool house; (DOI: 10-1098/rstb.2010.0387)

[10] International Journal of Computer Application (0975-8887). Improved J48 classification algorithm for prediction of diabetes.GaganjotKaur, AmitChabara.

[11]Korting, Thales Sehn, “C4.5 algorithm and multi variable decision tree” for National Institute of Space Research.

[12] Jones-McLean EM, Shatenstein B, Whiting SJ. Dietary patterns research and its applications to nutrition policy for the prevention of chronic disease among diverse North American populations. ApplPhysiolNutrMetab 2010; 35:195–198.

[13] U.S. Department of Health and Human Services and U.S. Department of Agriculture. Dietary Guidelines for Americans, 2010 [Internet]. Available from www .health.gov/dietaryguidelines/. Accessed 30 June 2011.

[14] Institute of Medicine. The Role of Nutrition in Maintaining Health in the Nation’s Elderly: Evaluating Coverage of Nutrition Services for the Medicare Population. Washington, DC, National Academies Press, 2000.

[15] Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P (2003) International physical activity questionnaire: 12-country reliability and validity. Medicine and Science in Sports and Exercise 35:1381–95.

[16] Blair SN, Haskell WL, Ho P, Paffenbarger RS Jr, Vranizan KM, Farquhar JW, Wood PD (1985) Assessment of habitual physical activity by a seven-day recall in a community survey and controlled experiments. Am J Epidemiol. 122:794–804.

[17] Ells LJ, Cavill N (2009) Preventing childhood obesity through lifestyle change interventions: A briefing paper for commissioners. Oxford: National Obesity Observatory.

[18]DariuszMatyja, “Application of data mining algorithms to analysis of medical data,” Master Software Engineering thesis, Blekinge Institute of Technology, Ronneby, Sweden, Aug. 2007.

[19]Jiawei Han and MichelineKamber, Data Mining: Concepts and Techniques,2nded.,Morgan Kaufmann Publishers., An Imprint of Elsevier, 2006.

[20]ZiemerDC, Berkowitz KJ, Panayioto RM, et al. A simple meal plan emphasizing healthy food choices is as effective as an exchange-based meal plan for urban African Americans with type 2 diabetes. Diabetes Care 2003;26:1719–1724.

[21] Magnus Stensmo and Terrence J. Sejnowski “Automated Medical Diagnosis based on Decision Theory and Learning from Cases “ World Congress on Neural Networks 1996 International Neural Networksociety pp. 1227-1231.