I J R B A T, Issue (IX), Vol. II, May 2021: 88-92

A Double-Blind Peer Reviewed & Refereed Journal

OPEN ACCES

e-ISSN 2347 – 517X

Original Article

Reg. No. : MH-NGP/00010

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY

© www.ijrbat.in

TEXT BASED FRUIT CLASSIFIER

T Sai Sravani, K Divya Teja and Sunil Bhutada
Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India
saisravani0427@gmail.com, sunilb@sreenidhi.edu.in and divya.t.katta@gmail.com

Communicated: 15.03.21	Revision : 21.04.2021 Accepted: 11.05.2021	Published: 30.05.2021

ABSTRACT:

Text classification is the method of analysing texts using a pre-processed dataset. Algorithms of machine learning generate a learning model and experimental results. The fruit dataset can be classified into different attributes. These results predict the popularity of the fruits and understand which machine learning Algorithm can give the highest accuracy for a particular dataset. In this paper, we perform classification algorithms on the fruit dataset and then develop a GUI to display the visualization of the results.

Keywords: Machine Learning, KNN, Fruit dataset, text classification

INTRODUCTION:

The amount of data only keeps increasing every day. This led to the development of technologies for instance artificial intelligence, cloud computing, and machine learning. The datasets are collected from various sources to analyse using different tools and algorithms. Text classification has several applications labelling, sentiment analysis, and anomaly detection. Text classification can be done manually yet, machine learning enhances the performance of surveys and accuracy. The three main approaches of text classification are

1) Rule-based

2) Machine learning

3) Hybrid Systems

Defining the category of text by a human or a simple code is considered a Rule-based method. The machine learning approach comprises of using statistical methods to perform classification, regression, and clustering. Combining the methods of rule-based and machine learning to analyse the text through Cross-Validation is categorized as Hybrid Systems. This gives a high accuracy and precision rate. The algorithms are run on the fruit dataset to generate estimated results and develop a GUI which is user-friendly and also provides the health benefits of a particular fruit.

LITERATURE REVIEW

After gathering information from the paper review to understand the process of Classification it has been mentioned in Z.Wang's paper that combining feature selection and modeling generates high accuracy experimental results. Pre-processing and Classification on text stream can help in eliminating the non relevant features from the data set. This way, the model developed has high performance. Also the author S. Di proposes the use of Hadoop in Natural language processing which aims to categorize the texts into different features. Evaluating the code with different models is important to understand which algorithm is feasible. Hadoop can also be used to perform text classification but due to the long queue of code, the data is clustered and increases the latency and processing time whereas the machine learning approach is scalable and quick to obtain the results. G.Aizhang states in his paper that determining the type of text using KNN Algorithm based on rough set can improve the efficiency and accuracy of the results even with a large amount of data. J.Chen has mentioned that Traditional Naive Byes doesn't perform well on small datasets; a co-relation factor is incorporated to achieve better results and accuracy.Text classification methods are surveyed to understand which algorithms generate high accuracy from predefined classes from a text document.KNN Algorithm is simple can be used further to improve the scope and precision of the datasets. The integration with Hadoop can also be done to take the advantage of distributed processing. Tools from Information retrieval can be used to search for data from web engines, databases, emails and articles for further performing multi-label classification.

PROPOSED SYSTEM

Step-1: Choosing a dataset

Datasets from the web, emails, articles contain missing values and noise, the pre-processing involves removal of these and make the dataset suitable to build a machine learning model.

Step-2: Training and Testing Data

First in order to generate accurate results the data is split into 70% training and 30% testing data to measure the goodness of fit of the model that is being developed.

Step-3: Machine learning tools to build a model

Algorithms like SVM, KNN, and Decision tree are run on the training dataset. After a model is developed it is applied on the testing data to get experimental results into an excel sheet.

Step-4: Metrics used for obtaining the results

As it takes both positive class predictions as well as negative, precision values are generated and f1-score to balance it out. Support values can further be used to generate confidence value.

Step-5: GUI for Query Options

Developing a GUI for representing the data in form of charts and adding other options like displaying health benefits and side effects of fruits from database.

Step-6: Experimental Results

A histogram with feature selection is displayed in the GUI and the Report can be viewed in Excel sheet.

Design Analysis

Generating an efficient learning model is the crucial part of the process. The aim to classify the data using different algorithms and identify the one that produces highest accuracy. After classification the results are visualized in a GUI for the users to understand the experimental results. These results can be integrated with the search engines and websites to understand what information the user is willing to fetch based on the organized and labeled data.

The entire process is only efficient when the right dataset and tools are used to train the model. K-Nearest Neighbour (KNN) Algorithm pertain to supervised learning technique. It observes the categorization of previous data and based on that assigns the category for the new case/data. The similarity between the prior dataset results and current one is the key factor for determining the class of new case/data.

METHODOLOGY:

Dataset

The dataset used for this paper is in the form of a CSV file (Excel Sheet) which is created by Dr. Iain Murray from University of Edinburgh. The dataset consists of different fruits and their physical attributes such as colour, mass, width, and height.

Implementation

After choosing the data set and pre-processing, we implemented classification Algorithms to generate training data set report. Then further applied the algorithms to testing data to calculate the accuracy of the models. The KNN Algorithm generated the highest accuracy for the particular data set. The result is visualized in the form of a Histogram in the GUI for any user to understand the content.

Experimental Result Analysis

Although the time complexity of KNN is high it gives accurate results. After implementing other algorithms, it is understood that size of data set and class labels are key factors for building a model that achieve maximum accuracy. The Support value can further be used to calculate confidence in order to determine the popular fruit and its features.

CONCLUSION

The process of Text classification has several and diverse applications. It can be integrated with the search engines, apps and improve the customer services by automation. The process is scalable and real-time analysis has been proved to be very efficient for marketing, estimate production sales and other business strategies. By building a GUI it enables domain suitability and display not just classification results but also health benefits and side effects of a particular fruit.

REFERENCES:

Z. Wang, J. Liu, G. Sun, J. Zhao, Z. Ding and X. Guan, "An Ensemble Classification Algorithm for Text Data Stream based on Feature Selection and Topic Model," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA),(2020)

Y. Zheng, "An Exploration on Text Classification with Classical Machine Learning Algorithm," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), (2019)

V. K. Vijayan, K. R. Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI),(2017)

Venkatesh and K. V. Ranjitha, "Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier," 2018 IEEE World Symposium on Communication Engineering (WSCE), (2018)

Nuipian, P. Meesad and P. Boonrawd, "A comparison between keywords and key-phrases in text categorization using feature section technique," 2011 Ninth International Conference on ICT and Knowledge Engineering, (2012)

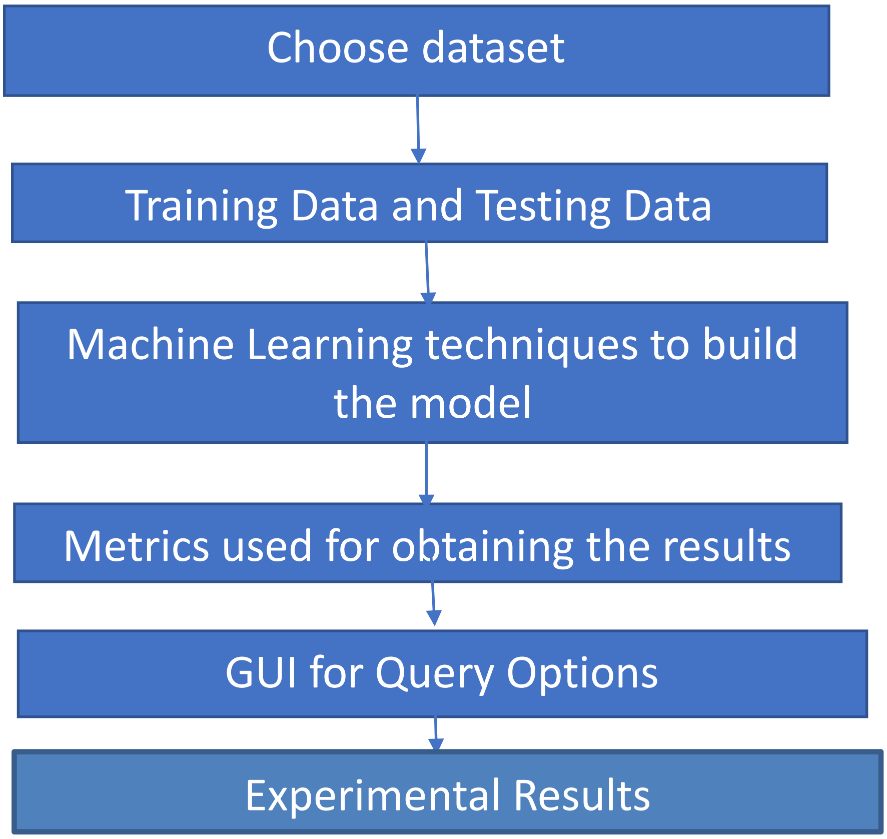
G. Aizhang and Y. Tao, "Based on Rough Sets and the Associated Analysis of KNN Text Classification Research," 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), (2015)

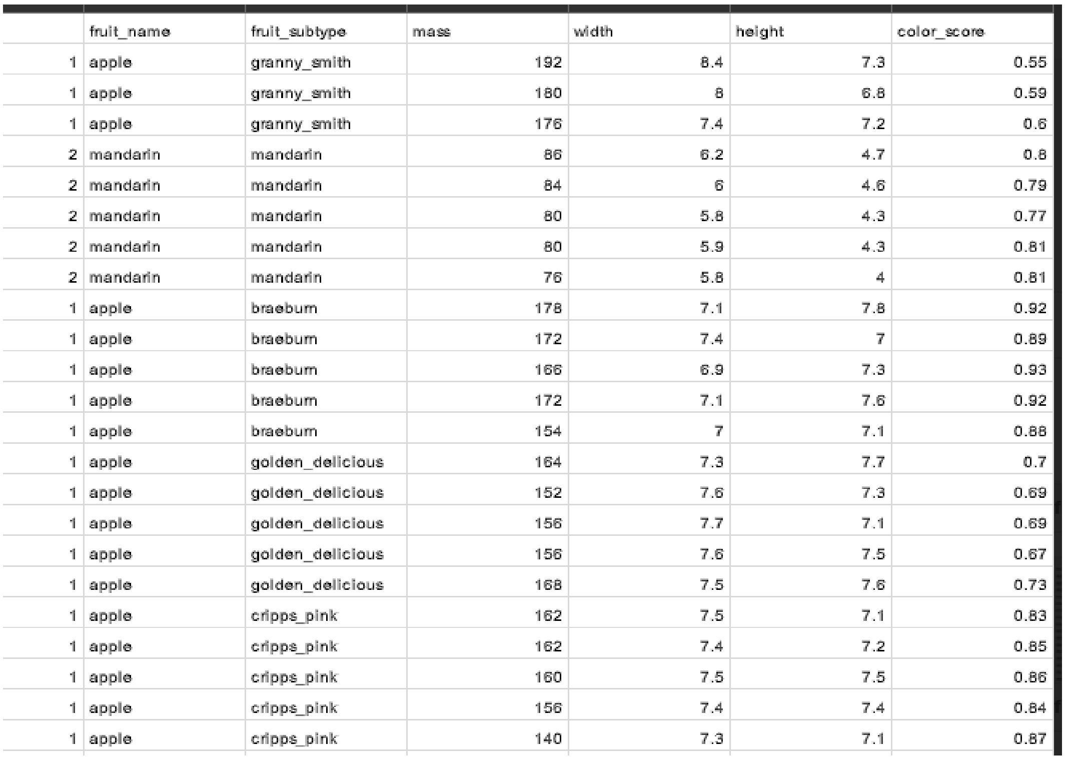
J. Chen, Z. Dai, J. Duan, H. Matzinger and I. Popescu, "Naive Bayes with Correlation Factor for Text Classification Problem," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA),(2019)

S. Du and J. Li, "Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop," 2019 7th International Conference on Information, Communication and Networks (ICICN),(2019)

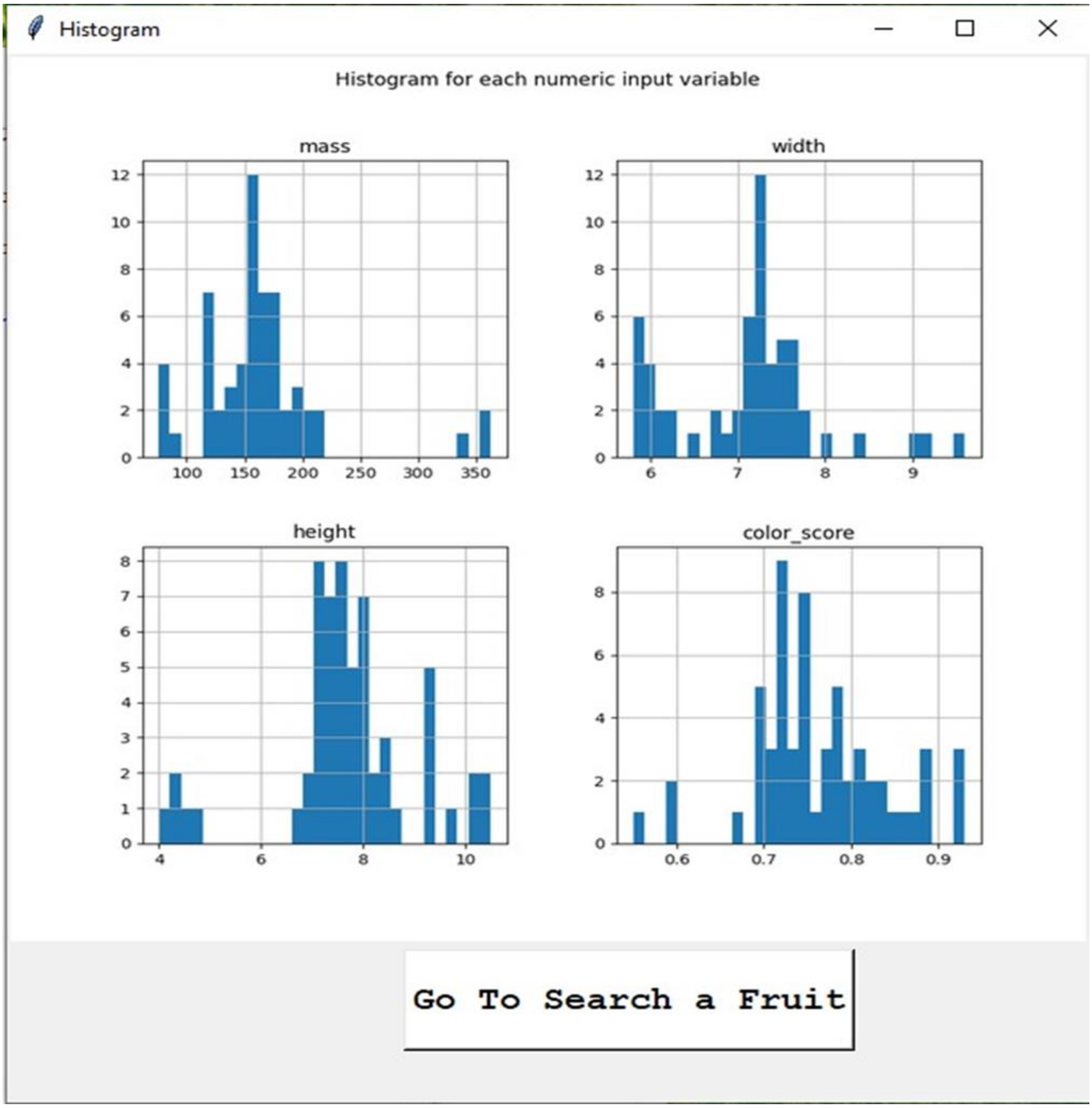
Y. Chang and H. Liu, "Semi-supervised classification algorithm based on the KNN," 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011

S. Lu, W. Tong and Z. Chen, "Implementation of the KNN algorithm based on Hadoop," 2015 International Conference on Smart and Sustainable City and Big Data (ICSSC)


Choose dataset
Training Data and Testing Data
Machine Learning techniques to build the model
Metrics used for obtaining the results
GUI for Query Options
Experimental Results



	fruit_name	fruit_subtype	mass	width	height	color_score
1	apple	granny_smith	192	8.4	7.3	0.55
1	apple	granny_smith	180	8	6.8	0.59
1	apple	granny_smith	176	7.4	7.2	0.6
2	mandarin	mandarin	86	6.2	4.7	0.8
2	mandarin	mandarin	84	6	4.6	0.79
2	mandarin	mandarin	80	5.8	4.3	0.77
2	mandarin	mandarin	80	5.9	4.3	0.81
2	mandarin	mandarin	76	5.8	4	0.81
1	apple	braeburn	178	7.1	7.8	0.92
1	apple	braeburn	172	7.4	7	0.89
1	apple	braeburn	166	6.9	7.3	0.93
1	apple	braeburn	172	7.1	7.6	0.92
1	apple	braeburn	154	7	7.1	0.88
1	apple	golden_delicious	164	7.3	7.7	0.7
1	apple	golden_delicious	152	7.6	7.3	0.69
1	apple	golden_delicious	156	7.7	7.1	0.69
1	apple	golden_delicious	156	7.6	7.5	0.67
1	apple	golden_delicious	168	7.5	7.6	0.73
1	apple	cripps_pink	162	7.5	7.1	0.83
1	apple	cripps_pink	162	7.4	7.2	0.85
1	apple	cripps_pink	160	7.5	7.5	0.86
1	apple	cripps_pink	156	7.4	7.4	0.84
1	apple	cripps_pink	140	7.3	7.1	0.87


Histogram
Histogram for each numeric input variable
mass
0
2
4
6
8
10
12
100
150
200
250
300
350
width
6
7
8
9
height
0
1
2
3
4
5
6
7
8
4
6
8
10
color_score
0.6
0.7
0.8
0.9
Go To Search a Fruit