MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Agricultural Research
ISSN: 2474-8846

Geospatial Distribution of the Quality of the Coffee of Minas Gerais by Methods of Data Mining Techniques

Barbosa JN[1]*, Sobrinho JPC[1], Silva TKB[1], Temoteo AS[2] and Bhering FP[1]

[1]Doctor in Plant Physiology, Federal Center for Technological Education of Minas Gerais, Brazil

[2]Professor of Statistics and Mathematics Federal Center for Technological Education of Minas Gerais, Brazil

***Corresponding author:** Juliana Neves Barbosa, Professor, Doctor in Plant Physiology, General Education Department, Federal Center for Technological Education of Minas Gerais, Brazil; Email: julianaanb@cefetmg.br, juliananevesbarbosa@gmail.com

Research Article
Volume 7 Issue 2
Received Date: June 20, 2022
Published Date: July 01, 2022
DOI: 10.23880/oajar-16000285

Abstract

The state of Minas Gerais is located in one of the most privileged Brazilian regions for the production of high-quality coffees. Due to its large territorial extension and environmental variation, it favors the production of specialty coffees with a great diversity of flavor and aroma. These differences are related to the peculiar characteristics of each microregion, particulary climate variations, altitude, latitude and production systems. Although the relationship between environmental characteristics and coffee productivity in different biomes is already well explored, its relationship with beverage quality still requires research. Given the need for more information on areas with potential for special coffee production, this paper proposes to associate data concerning to sensorial quality of coffees in the Cafés of Minas Quality Contest, with climatic and geographical variations of each state region, applying association rules for data mining, used to understand the aptitude of the state through the generation of rules that describe the most relevant patterns in the data. The climatic, geographical and classificatory information of the coffee samples is the result of the database of the contest participants in 2007 and 2008. The results demonstrate that with data mining techniques it was possible to associate the regions with the greatest potential with the climatic and geographical characteristics of the environment.

Keywords: Data Mining; Quality of Coffee; Mining Algorithm; Data Analytics

Introduction

Coffee production on Brazil has become one of the most prominent branches in the exports of agricultural products, due to the amount of bags produced annually and, mainly, to the quality presented [1]. This is due to the technological immersion on the field since the twentieth century, as well as the use of informational technology in the agribusiness management that made possible to add greater value and quality to the coffee produced in the country [2]. Minas Gerais is the state that has contributed the most to the highlight of coffee production, becoming the largest producer in the country, accounting for 78% of the national stocks and with a production of over 7 million bags of arabica coffee.

Minas Gerais also stands out for the production of high-quality coffees, awarded both nationally and internationally. Because it has favorable characteristics for coffee production in the northwest (NW), northeast (NE), midwest (MW), southwest (SW) and southeast (SE), and distinct in relation to climate and environmental factors among these regions, the state is capable of producing a variety of flavors with particularities of acidity, sweetness, body and aroma [3].

In addition to field-applied technologies, there are key factors in achieving excellence on coffee quality. Climate and altitude variables, for example, directly influences the developmental stages of coffee, causing distinct patterns of growth and development in the floral organs of the plant, which consequently generate a coffee differential in the national and international markets [4,5].

A prominent national competition that evaluates de quality of Minas Gerais coffees is the "Minas Gerais Quality Contest", which was created by a State Government's initiative, through the State Secretariat of Agriculture, Livestock and Supply, coordinated by the Company of Technical Assistance and Rural Extension of Minas Gerais (Emater-MG) and Federal University of Lavras (UFLA). The purpose of the event is to encourage the constant improvement of coffee quality to conquest markets, add value to the product and meet the growing demand for differentiated products. The identification of the best coffees in Minas Gerais is realized by BSCA (Brazil Specialty Coffee Association) tasters, who evaluate several aspects of the product, such as color, aroma, taste, body, acidity, among others. The contest comprises several stages and the coffees produced within the limits of the state are evaluated, being evaluated firstly the physical aspects and, in the later stage, the sensory attributes. The understanding of the environmental factors along with technological innovations can contribute with the increased production of specialty coffees in Minas Gerais and increasing quality improvement.

There are also competitions for research centers, such as the Brazilian Agricultural Research Corporation – EMBRAPA, which aims to encourage the development of new technologies in the field through the participation of universities and private companies.

To obtain a quality product in the field and understand the process that is favoring this production, it is necessary beyond the application of technology in the field and business management with the construction of a data generation workflow, the use of robust computational tools to analyze the information collected and turn it into a means of lowering costs, increasing the quality of your products and, consequently, business profit [6].

In the coffee industry, climatic aspects of the planting site may be closely linked to the quality of the beverage [7]. Due to the breadth of the regions of cultivation in the state, is necessary to analyze the combination of environmental and climatic factors (temperature, altitude, rainfall and humidity) to better understand their influence on coffee quality and contribute as a strategic tool for better use of these factors in the production of coffees with the higher quality potential [4,5].

The use of data mining techniques has proven to be a way of understanding and analyzing information from historical data. Data mining is a set of processing techniques for large databases aimed at extracting useful information through the discovery of previously unknown patterns, anomalies and relations in the data under study. To reach the level of information extraction, is necessary to go through several previous stages, such as data cleaning, data selection, data transformation, preprocessing, data mining and pattern assessment, all of which are of great importance to come up with real and relevant content results [8].

Therefore, for the analyses and evaluation of coffee quality in Minas Gerais, this work used the association rule, using an algorithm to associate the geographical and climatic factors with coffee quality.

Thus, the objective of this work was to analyze the quality distribution of coffees of Minas Gerais and to demonstrate the regions with the highest quality coffee production, applying data mining technique. For this, was used the contest's database from 2007 and 2008, which included the record sample scores and environmental factors of the participant cities. The results may contribute to future works that relate the potential of coffee production with quality, and consequently, promoting better strategies for coffee agribusiness.

Materials and Methods

To perform the data mining process, it is necessary to use techniques that extract information from the data and turn it into knowledge. For this, is used the process of Knowledge Discovery in Databases, which is divided into several stages, like data selection, preprocessing, transformation, mining and analysis and assimilation of results iteratively and with the goal to identify valid patterns and useful information from a data set [9-11].

Thus, in this section, we present the main stages of analysis and modeling using a custom data set that was made available through a partnership among the Federal University of Lavras (UFLA), Minas Gerais State Technical Assistance and Rural Extension Company (Emater-MG), Agricultural Research of Minas Gerais (EPAMIG), Brazilian Coffee Research Corporation (Embrapa Café) and the Federal Center of Technological Education of Minas Gerais (CEFET-MG), with data collected between 2007 and 2008.

The data used were from previous years to be able to understand the events that happened later without the use of any weather forecasting tool.

Preprocessing Step

In order to pre-process the data, data cleaning, data integration, data transformation and data reduction steps are performed, so it is possible to obtain an organized and generalized visualization of the data, thus preparing the data and choosing suitable techniques for the processing step [9,10].

Raw data: Contest evaluators' grade records are related to the environmental factors of each city sample. Only the attribute for the sample "Category" – natural or shelled – is categorical, the rest are all numeric.

Data organization: It was necessary to group the data by cities, because the environmental factors collected represented only the city of the samples, in this case, 1.161 samples were reduced to 117 participating cities. In the grouping, the arithmetic average of the scores obtained by each city was taken into account to provide sufficient information and make comparisons. The grouping was performed at three levels. The first one was the average between of the grades obtained in each stage of the contest, except the first stage that evaluates only physical aspects of the coffee. In the second grouping, it was calculated the average between the samples from the same city. And finally, an average was made between the two years in question.

Data reduction: At this stage, null and inconsistent data values are excluded, so that they don't interfere with the final result and creation of the developed predictive model. The categorization of quantitative data of environmental and climate variables into value ranges was also performed in order to improve the performance of the classification algorithm, as represented in Table 1.

Variables	Range of values
Temperature (ºC)	16,35~17
	17,1~20
	20,01~24,62
Altitude (m)	238,66~517,86
	517,87~912
	912,01~1248,78
Rain (mm)	867,58~1046
	1046,01~1300
	1300,01~1700,51
Moisture (%)	2,11~8,24
	8,25~49,7
	49,71~89,69
Regions	Northwest (NW)
	Northeast (NE)
	Midwest (MW)
	Southwest (SW)
	Southeast (SE)

Table 1: Value Range for categorization of quantitative data.

Choice of Algorithm

Then began the data mining step, to identify relationship patterns between attributes. In this step it is necessary to select the algorithm and configure the parameters of the association rule. In this work it was selected the Apriori algorithm, one of the most traditional mining algorithms using the frequent items strategy [12-20]. The Apriori algorithm works based on data association. It tells what possible combinations of a given A, with a given B could provide a better result. The differential of this algorithm is the use of the principle that if a set of items is frequent, their subsets will also be frequent. And if a subset has low frequency, their supersets will also have a low frequency. As algorithm works by combining the data, in the case of this work, coffee quality (contest grades) was related to climatic and geographic location data.

Algorithm parameters: The functioning of the algorithm is divided into two steps. In the first phase, it is necessary to establish the minimum support (relation of one variable with all variables at the same time) and in phase two, based on the minimum support, the association rules with an established minimum confidence factor are explored.

The support is calculated by dividing the number of transactions the record is present by the total transactions [16,17,20]. This is the amount that the transaction appears in the record, that is how many times we have the record in each interaction. Empirically, it was noted that the best support was 25%, as smaller values returned unwanted results and larger values did not return valid results. The general formula for support can be seen in equation (1).

$$\text{Support} = \frac{\text{Amount of transactions with the presented record}}{\text{Total Record Amount}} \quad (1)$$

The established trust can be calculated from the result of the amount that two records interact with the number of transactions of a record that is present in this transaction [13,18,19,20]. A hypothetical example would be how many times we have the southeast regions and an elevation of 1000 meters divided by the number of times the southeast region is present in all interactions. Based on this formula we could establish that the relationship between the records and, in practice, it was termed that it would be better if the confidence were 25%, however we can see a general formula for trust in equation (2).

$$\text{Thrust} = \frac{\text{Amount of transactions with the presented record}}{\text{Total Record Amount}} \quad (2)$$

Knowing that thrust and support values are the minimum values, all values found that are higher than them will be presented by the algorithm.

Training and Validation: As it is impractical to use all base records in the construction of the developed predictive model and to have a model independent of a specific data set, a sample of data was divided into training, validation and test data sets [17,18,20].

The training set is characterized by the grouping of records used in which the model is developed; the test set is the grouping used to test the constructed model; while the validation set refers to the set of records used to validate the constructed model. It is important to perform this division so that the model stays independent of another characteristic set, avoiding faulty results when submitted to other sets with different values from those used in the construction and validation of the model [9,10].

The models were developed in the validation and test data set. The following configuration was used for the models: data division 70% for model training and 15% for validation and testing to determine final model performance. This final division is one of the recommended standards for the creation of data mining and machine learning models, as well as presenting satisfactory results.

After development, the trainee models were then compared based on the error matrix and the curves for the validation dataset, in other words, we used the same models from the training set on the validation basis. Thus, for a satisfactory result, the models should present the same result.

Computational Tools

To implement the model, two computational tools were used. One of these is the Environmental for Knowledge Analysis (WEKA) tool an open-source tool in the GNU model, where you can edit and use the code as long as the creators are quoted, and it brings together a toolkit for data mining and machine learning. Used in early data analysis, the tool contains techniques implemented for data preparation, classification, regression, grouping, association rule mining and visualization [17,20,].

The Python programing language was also used, a free-use collaborative programming language that contains advanced libraries for data mining and machine learning, as well as data processing algorithms, linear algebra, among others [9,10,15,17,20]. The use of this programming language was to be able to validate the results obtained with the Weka too, that is, based on what was done and the results obtained in the tool it was possible to obtain and understand these same results with Python.

Results and Discussion

After performing all the steps, the algorithm was executed and then a model was generated that tells us the best combinations to get a good coffee quality rating. The variable that most influenced the coffee scores, in isolation, were those represented on Table 2.

Variables	Number of occurrences
Temperature (17.01°C~20ºC)	5
Altitude (517,87m~912m)	5
Rain (1300mm~1700,51mm)	4
Southwest region	4
Moisture (8,25%~49,7%)	4
Rain (1046,01mm~1300mm)	3
Temperature (20.01°C~24.62ºC)	3
Southeast region	2
Moisture (49,1%~89,69%)	2

Table 2: Variables that most contributed to the results.

The variables that most influenced the coffee scores, in isolation, were the temperature between 17,01ºC e 20ºC and altitude between 517,87 and 912 meters. But, in this case, with only one variable being evaluated, it is not possible to determine the quality of coffee produced in Minas Gerais. According to [16] environmental factors are decisive for the production of coffee quality. [5] reports that, faced with the need to identify properties that give flavor and aroma to coffees in the region of Minas Gerais, a tool is needed to ensure this quality standard. When combined, the pattern of environmental and climatic factors using data mining techniques is quite efficient in relation to the beverage quality of coffee produced in Minas Gerais. The result shown in Figure 1 represents the relationship between climate characteristics and coffee quality.

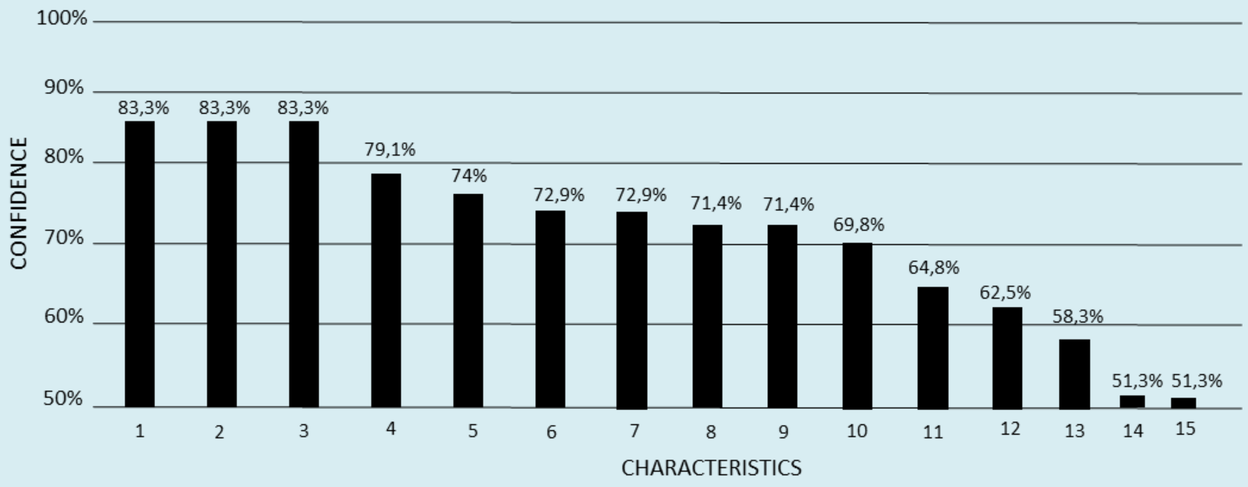

Figure 1: Results of relationships of climatic and geographic variables

The description of characteristics is in Table 3, along with the indexes presented in graphic 3.1. Values obtained for data confidence are medium values.

Characteristics	Subtitle	Confidence
1	Moisture Index (8,25% ~ 49,7%) and Rain (1046,01mm ~ 1300mm)	83,33%
2	Rain (1046,01mm ~ 1300mm) and Temperature (20.01°C ~ 24.62°C)	83,33%
3	Moisture Index (8,25% ~ 49,7%) and Rain (1046,01mm ~ 1300mm)	83,33%
4	Temperature (17.01°C ~20°C) and Rain (1046,01mm ~ 1300mm)	79,13%
5	Temperature (17.01°C ~20°C) and Moisture Index (49,71% ~ 89,69%)	74,00%
6	Temperature (17.01°C ~20°C) and Altitude (571,87m ~912m)	72,92%
7	Temperature (17.01°C ~20°C) and Southwest	72,92%
8	Southeast and Moisture Index (8,25% ~49,7%)	71,43%
9	Southeast and Temperature (20.01°C ~24.62°C)	71,42%
10	Altitude (517,87m ~912m) and Rain (1300mm ~1700,51mm)	69,83%
11	Altitude (517,87m ~912m) and Southwest	64,80%
12	Temperature (17.01°C ~20°C) and Altitude (517,87m ~912m)	62,50%
13	Altitude (517,87m ~912m) and Rain (1300mm ~1700,51mm)	58,33%
14	Moisture Index (49,71% ~ 89,69%) and Southwest	51,25%
15	Rain (1300mm ~1700,51mm) and Southwest	51,25%

Table 3: Description of the combination of characteristics shown in Figure 1.

According to the data analyzed in Figure 1 and Table 3, the Southeast region with temperature between 20,01ºC to 24ºC and humidity between 8,25% and 49,7% has a further 70% confidence in coffee harvesting with quality, such as the Southwest region combined temperature between 17,01ºC and 20ºC or humidity between 49,7% and 89,69% or altitude between 517,87 and 912 meters or rain between 1300mm to 1700,51mm. Some regions that are predominantly characterized by physical and geographical characteristics and specialized agricultural practice are the differentiating factors for a special production of special authors of special quality [3,4,14,15]. The large territorial extension and environmental variety that the state of Minas Gerais makes possible produces quality coffees with a great diversity of flavor and aroma. These differences are related to the particular characteristics of each municipality, mainly such as climatic variations, altitude and production systems [15]. Some climatic conditions were identified, such as the humidity index between 8,25% and 49,7% and rainfall between 1046.01mm and 1300mm annually, regardless of the region, as a confidence greater than 80% for coffee planting.

Conclusion

The article presented an analysis about the geospatial distribution of the quality coffees, demonstrating the regions with greater potential, showing that data mining is a valid technique for associating of the quality of coffees correlated with the climatic and geographical characteristics of the environment. According to the association rules of the Apriori algorithm, it was possible to observe patterns between coffee quality related to environmental characteristics. It was noteworthy that the data mining process can contribute to strategies and public policies for coffee agribusiness. The algorithm was executed on the Weka tool and validated with Python programming language with satisfactory results. For future work, we intend to expand the number of quantities researched and the application of other algorithms, aiming to search for patterns in the Coffee Quality Contest database that can, after analysis, be used in the future to estimate the potential for producing quality coffees based on its geographic location and climatic factors.

Acknowledgement: The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPq), the Brazilian Agricultural Research Corporation (EMBRAPA-CAFÉ) and the Minas Gerais State Research Support Foundation (FAPEMIG) and to the Federal Center for Technological Education of Minas Gerais (CEFET-MG) for the supporting the development of this work.

References

1. Mergulhão AD (2017) The flows, relationships and agents involved in the production and marketing of coffee currently produced in Brazil. Revista da Associação Nacional de Pós-graduação e Pesquisa em Geografia, 13(22): 57-85.

2. Redivo AR, Redivo A, Sornberger GP, Ferreira GA (2009) A Tecnologia De Informação Aplicada ao Agronegócio: Estudo sobre o sistema Agrogestor nas fazendas do Município de Sinop/MT. Revista Contabilidade e Amazônia 1(5).

3. Ramos ZU, Cesar CP, Marques FWP, Roberto CP (2017) Ambiente e variedades influenciam a qualidade de cafés das matas de minas. Coffee Science, Lavras 12(2): 240-247.

4. Barbosa JN, Borém FM, Cirillo MA, Malta MR, Alvarenga AA, et al. (2012) Coffee quality and its interactions with environment factors in Minas Gerais, Brazil. Journal of Agricultural Science 4(5): 181-190.

5. Barbosa JN, Borem FM, Alves HMR, Cirillo MA, Hanson C (2019) Assinatura isotópica da relação entre ambiente e qualidade espacial do café. African Journal of Agriculture Research 14: 354-360.

6. Horita FEA, Tanaka SA (2011) Business Intelligence (BI) aplicado em um Sistema Agropecuário. XVIII Simpósio de Iniciação Científica, Londrina.

7. Al Maolegi M, Arkok B (2014) An Improved Apriori Algorithm For Association Rules. International Journal on Natural Language Computing (IJNLC) 3(1): 21-29.

8. Fayyad U, Piatetsky Shapiro G, Smyth P (1996) From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence 17(3).

9. Camilo CO, Silva JC (2009) Data Mining: Concepts, Tasks, Methods and Tools. Universidade Federal de Goiás (UFC), pp: 1-29.

10. Destefani LA, Motyczka LB, Sausen P, Sausen A (2017) Busca de Padrões utilizando Algoritmo Bayesiano para Mineração de Dados nas Subestações Subterrâneas da CEEE. XXVIII Congresso Regional de Iniciação Científica e Tecnológica em Engenharia (CRICTE 2017), Ijuí, pp: 2318-2385.

11. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. 20th International Conference on Very Large Data Bases pp: 487-499.

12. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl 11(1): 10-18.

13. Lvov M, Kruglyk V (2014) Teaching Algorithmization and Programming using Python Language. Information Technologies in Education 20: 13-23.

14. Borem FM, Cirillo MA, Alves AP de C, Santos CM dos, Liska GR, et al. (2019) Coffee sensory quality study based on spatial distribution in the Mantiqueira mountain region of Brazil. Journal of Sensory Studies 35(2): 12552.

15. Silveira AS, Pinheiro ACT, Ferreira WPM, Silva LJ, Rufino JLS, et al. (2016) Sensory analysis of specialty coffee from different environmental conditions in the region of Matas de Minas, Minas Gerais, Brazil. Revista Ceres 63(4): 436-443.

16. Hailu BT, Maeda EE, Pellikka P, Pfeifer M (2015)

Identifying potential areas of understorey coffee in Ethiopia's highlands using predictive modelling. International Journal of Remote Sensing 36(11): 2898-2919.

17. Kawakubo FS, Machado RPP (2016) Mapping coffee crops in southeastern Brazil using spectral mixture analysis and data mining classification. Journal of Remote Sensing 37(14): 3414-3436.

18. Shankar SK, Kaur A (2016) Constraint data mining using apriori algorithm with AND operation. IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore pp: 1025-1029.

19. (2004) Weka 3: Machine Learning Software in Java. Disponível no site da University of Waikato.

20. Wang G, Yu X, Peng D, Cui Y, Li Q (2010) Research of data mining based on Apriori algorithm in cutting database. International Conference on Mechanic Automation and Control Engineering Wuhan pp: 3765-3768.