Salud, Ciencia y Tecnología - Serie de Conferencias. 2023; 2:568
doi: 10.56294/sctconf2023568

Categoría: Congreso Científico de la Fundación Salud, Ciencia y Tecnología 2023

ORIGINAL

Avocado Price Data Analysis Using Decision Tree

Análisis de datos sobre el precio del aguacate mediante un árbol de decisión

Johanes Fernandes Andry[1], Hendy Tannady[2] ✉, Glisina Dwinoor Rembulan[3], Aurellia Edinata[1]

[1]Information Systems Department. Universitas Bunda Mulia. Jakarta, Indonesia.
[2]Management Department. Universitas Multimedia Nusantara. Banten, Indonesia.
[3]Industrial Engineering Department. Universitas Bunda Mulia. Jakarta, Indonesia.

Citar como: Fernandes Andry J, Tannady H, Dwinoor Rembulan G, Edinata A. Avocado Price Data Analysis Using Decision Tree. Salud, Ciencia y Tecnología - Serie de Conferencias 2023; 2:568. https://doi.org/10.56294/sctconf2023568

Recibido: 19-06-2023 **Revisado:** 21-08-2023 **Aceptado:** 23-10-2023 **Publicado:** 24-10-2023

ABSTRACT

Data mining can be a method in which a great deal of information identifies useful patterns. The paper addresses some of the strategies of knowledge mining, algorithms and a few of the companies that have applied data processing technology to develop their businesses and have found outstanding results. Data Mining is the technique of digging and analyzing a very large volume of information in order to obtain something that is actual, new, very useful and can eventually find a style or pattern in the data. Data mining is a process of mining important information from data. The method of mining valuable knowledge from data is data mining. The Decision Tree is one way for Data Mining to forecast the future by constructing a model of classification or regression in the formof a tree structure. For exploring data into a decision tree that provides rules that are easy to understand in order to identify hidden relationships between input and target variables, the decision tree approach that transforms facts in the form of data is helpful. RapidMiner is one of the Mining Decision Tree Data Analysis methods. One of the aims of this analysis is to identify data on avocado prices dispersed in several stores in several American countries. The attributes used consist of the date of the sale, average price, total volume, avocado code, which have their respective meanings, such as: 4046, 4225 and 4770.

Key words: Data Mining; Decision Tree; Avocado Price; Rapid Miner.

RESUMEN

La minería de datos puede ser un método que permite identificar patrones útiles a partir de una gran cantidad de información. El documento aborda algunas de las estrategias de la minería del conocimiento, los algoritmos y algunas de las empresas que han aplicado la tecnología del procesamiento de datos para desarrollar sus negocios y han obtenido resultados sobresalientes. La minería de datos es la técnica de excavar y analizar un volumen muy grande de información con el fin de

obtener algo que sea actual, nuevo, muy útil y que, eventualmente, pueda encontrar un estilo o patrón en los datos. La minería de datos es un proceso de extracción de información importante a partir de datos. El método para extraer conocimientos valiosos de los datos es la minería de datos. El árbol de decisión es una forma de minería de datos para predecir el futuro mediante la construcción de un modelo de clasificación o regresión en la forma de una estructura de árbol. Para explorar los datos en un árbol de decisión que proporciona reglas que son fáciles de entender con el fin de identificar las relaciones ocultas entre las variables de entrada y de destino, el enfoque de árbol de decisión que transforma los hechos en forma de datos es útil. RapidMiner es uno de los métodos de análisis de datos en forma de árbol de decisión. Uno de los objetivos de este análisis es identificar datos sobre los precios del aguacate dispersos en varias tiendas de varios países americanos. Los atributos utilizados consisten en la fecha de la venta, precio medio, volumen total, código del aguacate, que tienen sus respectivos significados, tales como: 4046, 4225 y 4770.

Palabras clave: Minería de Datos; Árbol de Decisión; Precio del Aguacate; Rapid Miner.

INTRODUCTION

Persea americana, generally referred to as avocado, first appeared thousands of years ago in Mexico, but it was not introduced to California, USA, until 1871. By the 1950s, hundreds of varieties were available in the country's markets, with the variety most consumed being Fuerte. This condition changed about twenty years later, and the Hass avocado began to be the most consumed variety in the nation and worldwide. Avocado consumption currently exists not only because of its taste, but also because of its balanced contribution to people's diets.[1]

At present, 85 percent of the world's avocados produced and sold are of the Hass variety. This variety grows virtually all year round and in various regions. Mexico, the United States, Chile, Australia, South Africa and Israel are some of the leading producing nations, with Mexico being the world's largest producer, accounting for around a third of production worldwide.[2] Since 2008, the avocado market in the United States has risen 16 percent every year and this trend is expected to continue, at least in the medium term. The avocado producers in this country are states such as Florida, California and Hawaii, but the production does not reach consumer demands, so avocados are imported from, among other nations, Mexico, Chile, Peru, New Zealand and the Dominican Republic. The consumption of avocado is not standardized across the region, however. for example, about ninety% of the households in California devour avocados, in a share of extra than 3 gadgets regular with month. however, in some states of the outstanding Plains, first-rate a bit greater than half of of the households devour this fruit, in a percentage of no extra than devices steady with month.[3]

Accumulating statistics available on the market and on better practices regarding avocado cultivation would be of fantastic help to producers, carriers, associations, and organizations. this can be used to pick out the proper places to sell avocados, to carry out a hit advertising and marketing campaigns or to increase innovations for the production and income of such product. An example of this fact is that approximately fifty million greenbacks are spent yearly on advertising and marketing and promotional sports on healthy avocado consumption.[3] several authors have discovered that weather is one of the elements affecting financial and business conduct.[4,5,6,7]

There may be a department of synthetic intelligence referred to as machine gaining knowledge of to are expecting destiny moves or relationships, which entails constructing class and regression models,[8] for

example, it's miles viable to are expecting crop manufacturing[9] and disorder incidence[10,11] by means of the use of several machine gaining knowledge of techniques and also by means of amassing weather data. using weather and marketplace data, fashions used to expect avocado income in the u.s. can be evolved. It will be of great benefit to manufacturers, retailers, associations and businesses to collect information on the demand and on best practices regarding avocado cultivation. This could be used to select the best locations to sell avocados, to execute effective marketing campaigns or to create creative products for the manufacture and sale of those products. An example of this fact is that about $50 million is spent on ads and promotional activities on safe consumption of avocado every year.

Literature review

In this study, to collect avocado price data to process it in rapid mining, the following methods are needed, namely:

Data Mining

Data mining is the method of selecting, exploring and modeling vast volumes of data to define regularities or relationships that are initially unknown in order to produce consistent and useful consequences for the database owner. statistics mining is a beneficial device for strategic decision-making and performs an crucial function in marketplace segmentation, purchaser offerings, identity of fraud, credit and movements.[12] Data mining can generally be identified in large quantities of information as a technique for locating patterns (extraction) or interesting records. This method has been broadly utilized in research in areas consisting of health, engineering, marketing, education, and others. Data mining techniques are used to evaluate the relationship between subjects in the curriculum within the field of education.[13] The evolution of data mining starts with the first data collection times, in particular for business applications and bioinformatics applications stored in the computer, and the rise in data access technology.[14] Data mining is a market tool for exploring vast volumes of knowledge in order to discover meaningful trends and rules.[15] Including: clustering, grouping, estimation and association,[16] there are many data mining techniques.

Clustering

Clustering is a technique divided records to the cluster based at the similarity. ok-manner is a cluster method facts that the point information of lifestyles in cluster is rely upon diploma of the member.[17] In the method of clustering the principle idea is emphasized is the valuable seek cluster iteratively, in which the middle decided primarily based at the minimum distance of each facts at the center cluster.[18]

Rapid Miner

RapidMiner is one of the tools used for data mining to analyze web-accessed information. It is used for science, education, fast prototyping, development of software, and industrial applications. It is a licensed Open Source program that involves data cleaning, transformation of data, optimization, validation and visualization. The visualization consists of viewing the analyzed facts within the form of scatter plot, Bar, Pie chart, etc. additionally it is the various clustering and class algorithms to do the analytical process. one of the most important feature of this device that, it's going to analyze data with none application coding, however if anybody wants to analyze the records with their own coding then it may additionally include in the device.[19]

https://doi.org/10.56294/sctconf2023568

METHODS

Data Collection

After using the appropriate data mining methods. This researcher collected data in the form of avocado price data and took sample data on the Kaggle website.[20]

The avocado price data obtained must go through pre-processing before getting the best value when the program execution process takes place in rapid miner. Data cleaning and cutting will be done in this phase to repair some corrupted data. The data received is not good because a lot of data is empty and does not require attributes. After that the data is transformed into the appropriate data so that it can be processed using a decision tree based algorithm. so that the asset data is obtained for processing to the next stage.[21]

Data Mining

Information Mining is the procedure of extracting statistics from statistics units thru the use of algorithms and techniques involving the fields of records, gadget studying, and database control structures. In this study, researchers used a decision tree based algorithm. This researcher uses this method to make decisions in estimating a case, then the researcher analyzes the data such as where the data is obtained and what additional attributes are needed.[22]

Analysis

The analysis degree is carried out to attain the avocado price information grouping pattern. The device used is the rapid Miner. This device is extensively utilized in statistics technology, including for records coaching, system mastering, textual content mining, and predictive analysis.[23]

Validation

In the data validation stage, the avocado price is checked against the previously determined finding documents. This stage aims to ensure that the document findings submitted by the participants are accurate.

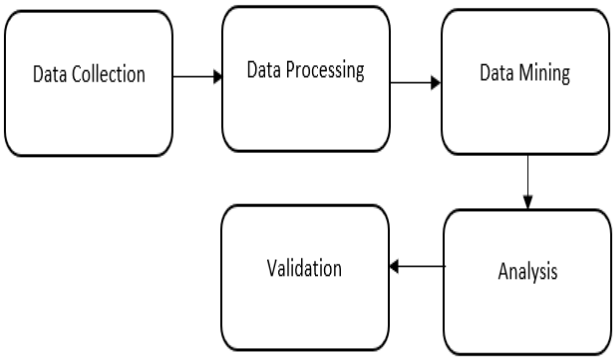

Figure 1. Research of Method[23]

DISCUSSION

Before using the Rapid Miner application, first make sure the data will be processed. The data selected is avocado price data in supermarkets domiciled in America with a period of 2015 in some areas such as Atlanta, Albany, etc. When all data is valid then start processing data using RapidMiner Studio application.

https://doi.org/10.56294/sctconf2023568

Dataset

"In this study, the dataset to be used comes from Kaggle with the name" Avocado prices. From 2015 to 2018, this dataset gathered 18249 data. According to the Kaggle website, this dataset mainly contains numbers and texts without any missing or damaged info.

In table 1, we will look at the types of Kaggle data in the table below, such as the sales of any city, the yearly sales, the form sold, and the average price of the avocado sales proceeds. Table 1 that contain list of attributes that used In this dataset :

Table 1. Attributes of the dataset

Attributes	Description
Date	The date of observation and taken record.
Average price	The average price of a single avocado in said region
Total volume	Total number of avocados sold in said region
Type A	Total number of avocados with PLU 4046 sold
Type B	Total number of avocados with PLU 4225 sold
Type C	Total number of avocados with PLU 4770 sold
Total bags	Total of all bags used for avocado purchased in said region
Small bags	Total small bags used
Large bags	Total large bags used
X Large bags	Total XL bags used
Type	can be Conventional or organic.
Year	The year the data recorded.
Region	The city or region of the observation being taken. A total of 52 region used.

Data Processing

Data Processing is the first step to take. Here, the data from any suspected lost or broken data will be cleaned up. This method is performed using the "turbo prep" data cleaner in rapid miner.

Next, importing our data to the quick miner is the first thing to do. It is important to check the existence of our data attributes when importing our data.

There are several attribute types that is used in this research:

1. Polynomial: Type of data where the data contain more than possible answer.
2. Binominal: type of data where it only contains two answers, YES (1) or NO (0).
3. Real: type of data where the number contain decimal.
4. Integer: type of data where the number is a fixed number without decimal.
5. Date: type of data that representing Date

Make sure each attributes have the correct type as such bellow:

1. Date: Date
2. Average price: Real
3. Total volume: Real
4. Type A: Real
5. Type B: Real
6. Type C: Real
7. Total bags: Real

https://doi.org/10.56294/sctconf2023568

8. Small bags: Real
9. Large bags: Real
10. X Large bags: Real
11. Type: Binominal
12. Year: Integer
13. Region: Polynomial (Label)

Data with the wrong attribute may or may not split the outcome or even cause an error in the calculation process.

After ensuring that each attribute has the correct form, we can proceed to the next stage, choosing the available turbo prep right after we have imported the data.

It was proven correct after review using the method that the data did not contain any errors or missing data in the dataset.

Decision tree

After data processing is done, the next step is the Implementation of decision tree on the data in rapid miner. This decision tree will help us in finding the region where most avocados sold every year.

First step is, we'll be importing our avocado data into the main field. after that, we'll want to add operator "decision tree" to the field, to actually doing the decision tree calculation. Then we'll be connecting the avocado data to the decision tree operator via the "out" section of the avocado data to the "tra" section of the decision tree. After that, we'll be connecting the "mod" section of the decision tree to the "res" on the rightmost of the field.

The depth that used on the decision tree is 7, for the sake of simplicity. If done right, the next step is to "run" the said model using "play" button to let the machine calculating and processing all the data. After the process is done, we'll be shown the decision tree that it has created as shown In figure 2.

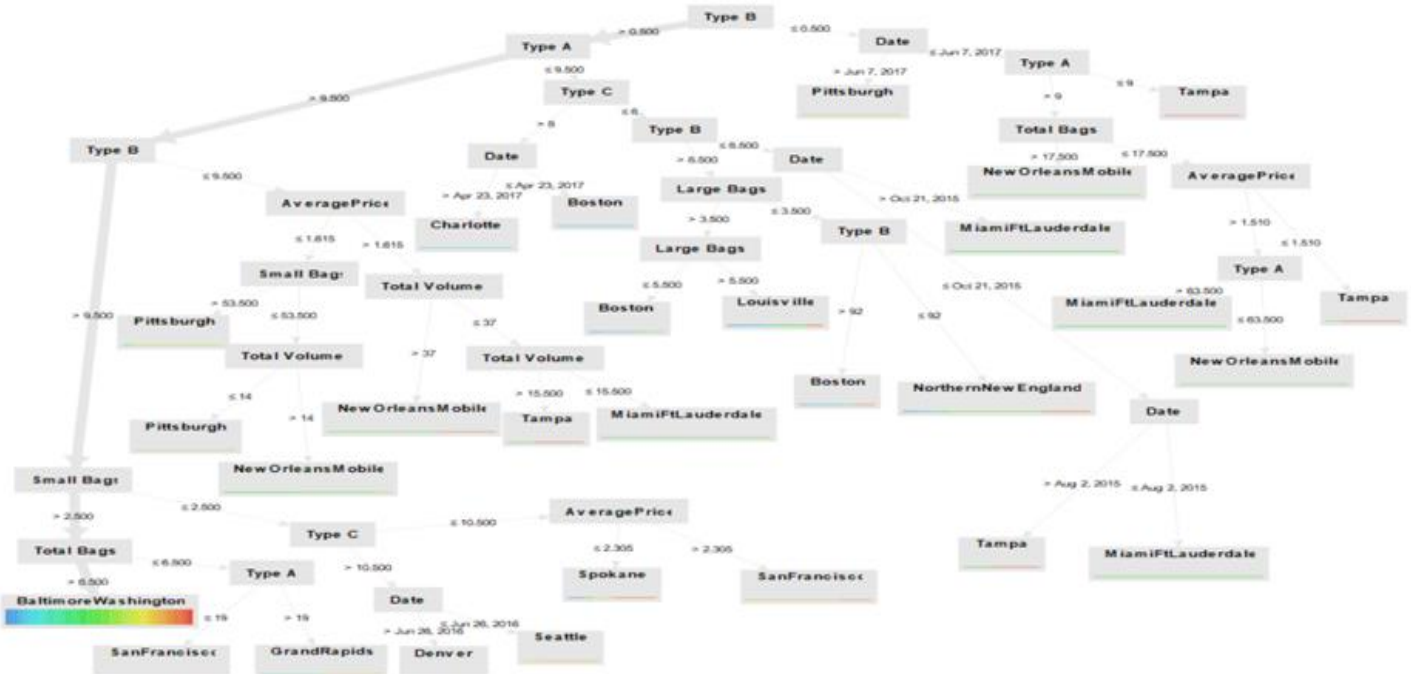

Figure 2. Decision Tree Result

https://doi.org/10.56294/sctconf2023568

The decision tree In Figure 3 has been created can be used to predict which region can or will sell the most avocado in the coming future and also searching for the current trending type of avocado being sold there.

Diagram

In case the decision tree is hard to understand, we can also see the result in diagram form. The diagram that will be used is line diagram. The data that being used is still the same avocado data, without any modification the data.

First step is, we'll be importing our avocado data into the main field. after that, add operator "split Data". connect the avocado dataset into the split data. The split data is used to splitting the data into several pieces, in this case it will splice the data into decision tree usage and apply model usage. For the decision tree, the split percentage that will be used is 10 % of the data, and for the apply model is 90 %.

The next step is inserting decision tree operator into the field. Connect the operator split data into the decision tree. Make sure the split that connected is the 10 % one. This data will be used by the operator as benchmark for the main data.

The last step is inserting the apply model operator. Connect both the split data (90 %) and the decision tree operator into the apply model, and then connect the apply model into "res". If the process done right, it should become like figure 3.

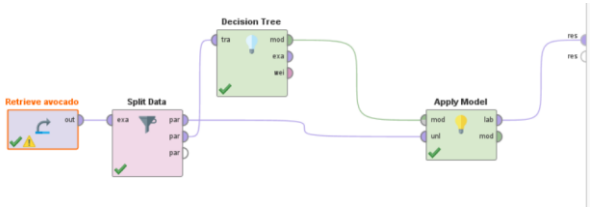

Figure 3. Final model of the decision tree

If done right, the next step is to "run" the said model using "play" button to let the machine calculating and processing all the data. After the calculation done, we'll be shown the result. Head to statistic and we'll be shown the confidence of each region (54 region) in detail in Figure 4, including min and max of the confidence value.

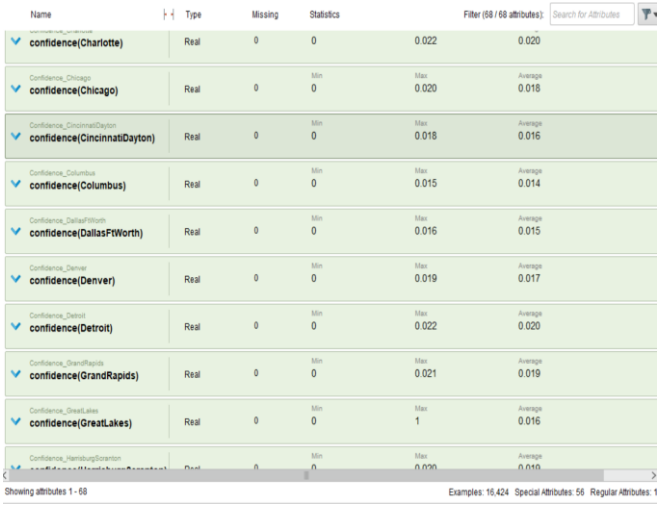

Figure 4. Confidence statistic in each region

We can also head to visualization to see the graph result using certain values. In this case, we'll be examine the graph of total avocado being sold in correlation to the price per individual avocados.

In figure 5, we'll be able to examine where region sold avocados the most. In that chart, we'll be able to take conclusion that Boise region sold the most avocados by significant amount compared to other region by 60M+ of

https://doi.org/10.56294/sctconf2023568

avocados being sold. Also, upon examining the graph for sales per prices, its determined that the best price for the avocado sales is $0,87.

Upon conclusion, if we were to choose the best place to sold region at certain price, if we want the best result, we will choose to market it at Boise at price around $0,80 ~ $0,90 for the best result in sales. The result may change along with times, so frequent analysis and data collection must be done frequently or periodically in order to get the best result each time we want to choose the best market strategy possible.

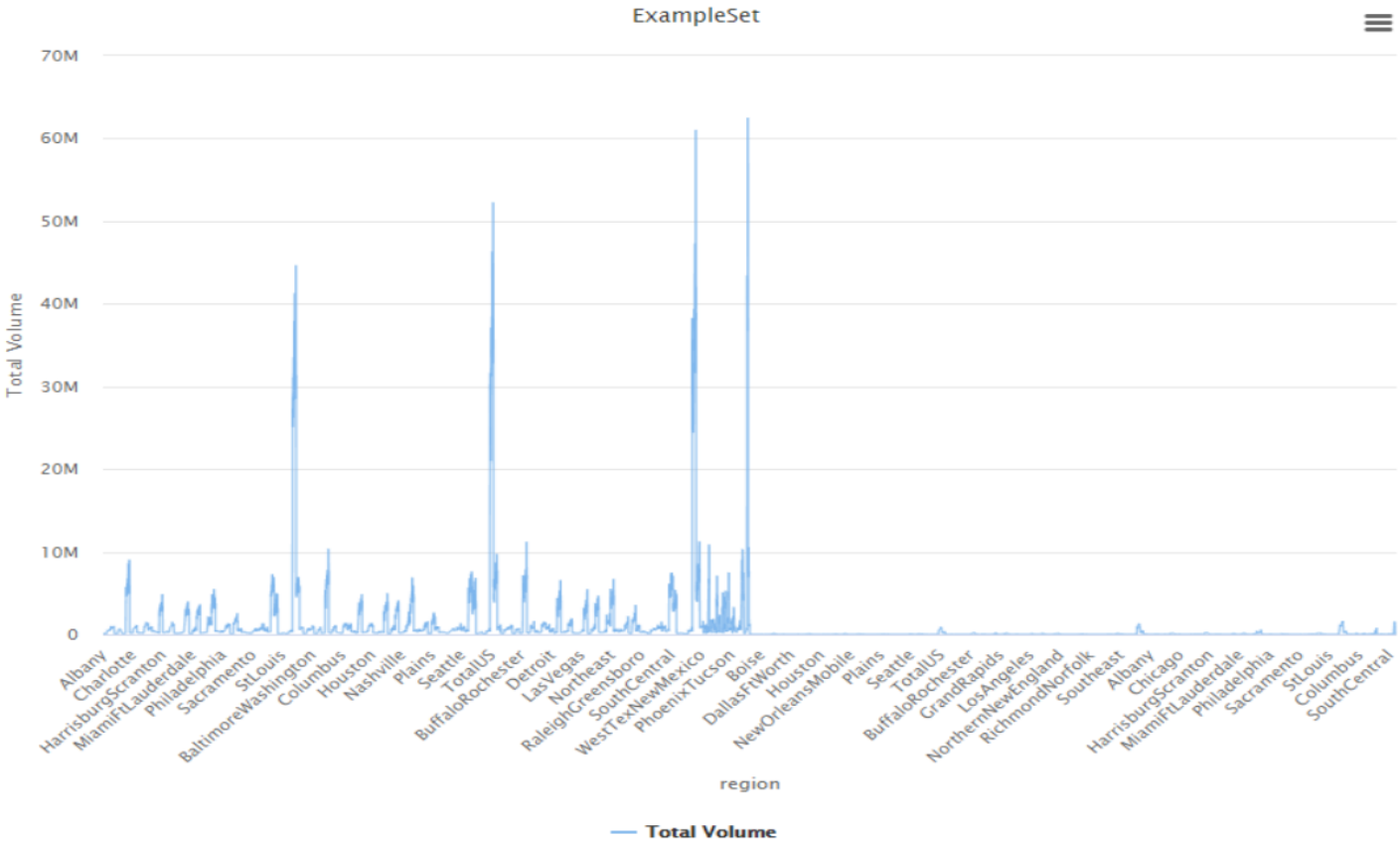

Figure 5. Total avocados being sold in regions

CONCLUSION

Based on the previously stated description, it can be concluded that the Decision Tree Method can be used to group the sales data of avocados scattered in several regions, from the most sold, sold, and under-sold. The Decision Tree method can also be used to assist the store in grouping accessories, based on the results of manual comparison with the software there are differences in results. If manually all data looks very much and complicated, when done using the Decision Tree Method all the data becomes look neater. Another advantage of using the Decision Tree Method is that it can explore records, finding hidden relationships among a number of potential enter variables and a goal variable. decision Tree combines information discover and modeling, so it could be properly as a primary step in the modeling manner even when used as the very last version of some different techniques. In a few programs, the accuracy of a category or prediction is the only aspect highlighted on this technique, but it has the ability to eliminate records that might otherwise be no longer essential. due to the fact, existing samples are normally simplest tested based on precise standards or instructions just the same as when the authors conducted a study on avocado sales data.

The author realizes there are still flaws in this writing, due to the limitations of the author in terms of knowledge. In order to correct the shortcomings and to improve this research the authors provide some suggestions for further research can be done with more data and more parameters as well as to maximize process time.

https://doi.org/10.56294/sctconf2023568

SUGGESTION

In this study there are still many shortcomings that still need to be developed, then if the lack of this research can be corrected in the next research.

To keep this research growing, here are the proposed suggestions:

1. This research can be developed with other classification data mining methods to make comparisons.
2. It is expected to analyze more about avocados that affect the level of sales of avocados so that by analyzing, sellers can know the mistakes they have made, and can increase those mistakes so as not to repeat them.

REFERENCES

1. D. M.L, Dreher and A.J, “Has Avocado Composition and Potential Health Effects,” Crit. Rev. Foof Sci. Nutr., vol. 53, pp. 738–750, 2013.

2. Silva, “Avocado History, Biodiversity and Production In Sustainable Horticultural Systems,” Springer Int. Publ., pp. 157–205, 2014.

3. G. Cavaletto, “The avocado market in the United States In Proceedings of the VIII Congreso Mundial de la Palta,” pp. 463–466, 2015.

4. K. and Y. Furuya, “Impacts of climate change on rice market and production capacity in the Lower Mekong Basin.,” Paddy Water Env., vol. 12, pp. 255–274, 2014.

5. Y. Kang, Jiang, Zee, “Weather Effects on the Returns and Volatility of the Shanghai Stock Market.,” Phys. A Stat. Mech. Appl., vol. 389, pp. 91–99, 2010.

6. Murray and Di Muro, “The Effect of Weather on Consumer Spending.,” J. Retail. Consum. Serv, vol. 17, pp. 512–520, 2010.

7. S. L, “Does The Weather Affect Stock Market Volatility?,” Financ. Res. Lett., vol. 7, pp. 214–223, 2010.

8. J. F. Andry, H. Tannady, I. I. Limawal, G. D. Rembulan and R. F. Marta, "Big Data Analysis On Youtube With Tableau," Journal of Theoretical and Applied Information Technology, vol. 99, no. 22, pp. 1915-1929, 2021.

9. C. Plazas, Lopez, “A Tool for Classification of Cacao Production in Colombia Based on Multiple Classifier Systems.,” Springer Int. Publ., pp. 60–69, 2017.

10. Corrales and Casas, “Two-Level Classifier Ensembles for Coffee Rust Estimation in Colombian Crops.,” Int. J. Agric. Environ. Inf. Syst., vol. 7, pp. 41–59, 2016.

11. Marcelo KVG, Claudio BAM, Ruiz JAZ. Impact of Work Motivation on service advisors of a public institution in North Lima. Southern Perspective / Perspectiva Austral 2023;1:11–11. https://doi.org/10.56294/pa202311.

12. David BGM, Ruiz ZRZ, Claudio BAM. Transportation management and distribution of goods in a transportation company in the department of Ancash. Southern Perspective / Perspectiva Austral 2023;1:4–4. https://doi.org/10.56294/pa20234.

13. C. Lasso, Valencia, “Decision Support System for Coffee Rust Control Based on Expert Knowledge and Value-Added Services.,” Springer Int. Publ., pp. 70–83, 2017.

14. Keramati and N. Yousefi, “A Proposed Classification of Data Mining Techniques in Credit Scoring,” Techniques, pp. 416–424, 2011.

15. Harwati, A. P. Alfiani, and F. A. Wulandari, “Mapping Student’s Performance Based on Data Mining Approach (A Case Study),” Agric. Agric. Sci. Procedia, vol. 3, pp. 173–177, 2015, doi: 10.1016/j.aaspro.2015.01.034.

16. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, “Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm,” Int. J. Eng. Emerg. Technol., vol. 2, no. 1, p. 53, 2017, doi: 10.24843/ijeet.2017.v02.i01.p11.

17. Dionicio RJA, Serna YPO, Claudio BAM, Ruiz JAZ. Sales processes of the consultants of a company in the bakery industry. Southern Perspective / Perspectiva Austral 2023;1:2–2. https://doi.org/10.56294/pa20232.

18. Velásquez AA, Gómez JAY, Claudio BAM, Ruiz JAZ. Soft skills and the labor market insertion of students in the last cycles of administration at a university in northern Lima. Southern Perspective / Perspectiva Austral 2024;2:21–21. https://doi.org/10.56294/pa202421.

19. W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, “Supervised data mining approach for predicting student performance,” Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, pp. 1584–1592, 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

20. R. Rahim, “Educational Data Mining (EDM) on the use of the internet in the world of Indonesian education,” TEM J., vol. 9, no. 3, pp. 1134–1140, 2020, doi: 10.18421/TEM93-39.

21. W. Purba, S. Tamba, and J. Saragih, “The effect of mining data k-means clustering toward students profile model drop out potential,” J. Phys. Conf. Ser., vol. 1007, no. 1, 2018, doi: 10.1088/1742-6596/1007/1/012049.

22. Y. Sinambela, S. Herman, A. Takwim, and S. R. Widianto, “a Study of Comparing Conceptual and Performance of K-Means and Fuzzy C Means Algorithms (Clustering Method of Data Mining) of Consumer Segmentation,” J. Ris. Inform., vol. 2, no. 2, pp. 49–54, 2020, doi: 10.34288/jri.v2i2.116.

23. Ñope EMG, Claudio BAM, Ruiz JAZ. The Service Quality of a Feed Industry Company. Southern Perspective / Perspectiva Austral 2023;1:9–9. https://doi.org/10.56294/pa20239.

24. Jeronimo CJC, Basilio AYP, Claudio BAM, Ruiz JAZ. Human talent management and the work performance of employees in a textile company in Comas. Southern Perspective / Perspectiva Austral 2023;1:5–5. https://doi.org/10.56294/pa20235.

25. M. Santhanakumar and C. C. Columbus, “Web Usage Based Analysis of Web Pages Using RapidMiner Department of Computer Science and Engineering,” WSEAS Trans. Comput., vol. 14, pp. 455–464, 2015, Online]. Available: http://www.wseas.org/multimedia/journals/computers/2015/a925705-700.pdf.

26. E. D. Madyatmadja, J. F. Andry, and A. Chandra, “Blueprint enterprise architecture in distribution company using togaf,” J. Theor. Appl. Inf. Technol., vol. 98, no. 12, pp. 2006–2016, 2020.

27. T. Sitorus and H. Tannady, "Synergy, System IT, Risk Management and The Influence on Cyber Terrorism and Hoax News Action", Journal of Theoretical and Applied Information Technology, vol. 99, no. 8, pp. 1802-1814, 2021.

28. Anton, “Data Mining Implementation with Clustering Techniques for Drug Inventory Information in Antonius Hospital Pontianak,” Eduma Math. Educ. Learn. Teach., vol. 9, no. 1, p. 86, 2020, doi: 10.24235/eduma.v9i1.4011.

29. N. Uddin, H. Hermawan, N. L. Rachmawatia and H. Tannady, "Genetic Algorithm for Logistics-Route Optimization in Urban Area", 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp. 071-076.

FINANCING

The authors did not receive funding for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Johanes Fernandes Andry, Hendy Tannady, Glisina Dwinoor Rembulan, Aurellia Edinata.

Research: Johanes Fernandes Andry, Hendy Tannady, Glisina Dwinoor Rembulan, Aurellia Edinata.

Writing-original draft: Johanes Fernandes Andry, Hendy Tannady, Glisina Dwinoor Rembulan, Aurellia Edinata.

Writing-review and proof editing: Johanes Fernandes Andry, Hendy Tannady, Glisina Dwinoor Rembulan, Aurellia Edinata.