WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# BigMart Sales Prediction: Insights from Data Science

***[1]Dr Kondragunta Rama Krishnaiah , [2]Dr Muvva Venkateswara Rao***

[1]Professor, Dept. of CSE, R K College of Engineering, Vijayawada-521456, A.P, India.
[2]Professor, Dept. of CSE, NRI Institute of Technology, Vijayawada- 521212, India

**ABSTRACT**

Nowadays shopping malls and Big Marts keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. For each customer we know what the individual products (items) are that he has put in his basket and bought. Other use cases for MBA could be web click data, log files, and even questionnaires

With market basket analysis we can identify items that are frequently bought together. Usually the results of an MBA are presented in the form of rules. The rules can be as simple as {A ==> B}, when a customer buys item A then it is (very) likely that the customer buys item B. More complex rules are also possible {A, B ==> D, F}, when a customer buys items A and B then it is likely that he buys items D and F

We propose a predictive model using Decision tree regression technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. A comparative analysis of the model with others in terms performance metrics

**KEYWORDS:** machine learning, Sales forecast, sales prediction, Pandas, Numpy, linear regression, ridge regression, Decision tree Regression,

## 1. INTRODUCTION:

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems.

We are addressing the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc. are used for prediction or forecasting of sales volume.

### *1.1 DOMAIN:*

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and in sights .from many structural and unstructured data. Data science is related to data mining and big data.

### *1.2 MACHINE LEARNING:*

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

### *1.3 BACKGROUND*

We are using pandas for handing data and numpy for handling numerical operations in arrays.

**Pandas**

Python has long been great for data munging and preparation, but less so for data analysis and modeling. pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R. Combined with the excellent

IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate. pandas does not implement significant modeling functionality outside of linear and panel regression; for this, look to stats models and scikitlearn. More work is still needed to make Python a first class statistical modeling environment, but we are well on our way toward that goal.

**NumPy**

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code

Useful linear algebra, Fourier transforms, and random number capabilities. Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data.

## 2. LITRATURE REVIEW:

Journals and papers were studied which relates the content on sales forecast prediction using machine learning algorithms. Below are the list of few papers which were studied and a review of the paper was added along.

Due to the strong competition that exists today, most manufacturing organizations are in a continuous effort for increasing their profits and reducing their costs. Accurate sales forecasting is certainly an inexpensive way to meet the a fore mentioned goals, since this leads to improved customer service, reduced lost sales and product returns and more efficient production planning.

Especially for the food industry, successful sales forecasting systems can be very beneficial, due to the short shelf-life of many food products and the importance of the product quality which is closely related to human health.

Association rules (frequent item sets), classification and clustering are main methods used in data mining research. One of the great challenges of data mining is finding hidden patterns without violating data owners' privacy. Privacy preserving data mining came into prominence as a solution. In the aim of the paper. Different prediction methods give different performance predictions when used for daily fresh food sales forecasting.

**Limitations:**

- Limited Number of Combinations.
- When Data Size Increase It Become Hard To Analyze Data Set.
- It Is Not User Friendly.
- Less Effective
- Eventually Problem Size Increase
- Limited Number of Graphical Libraries

## 3. Research Methodology:

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps as data set, data exploration, data cleaning, feature engineering, model building, model testing. we propose a model using decision tree regression Technique. Every step plays a vital role for building the proposed model. In our model we have used 2013 bigmart dataset. After preprocessing and filling missing.

.Values, we used ensemble classifier using Decision trees, Linear regression, Ridge regression and decision tree regression. Both MAE and RSME are used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum MAE and RSM.

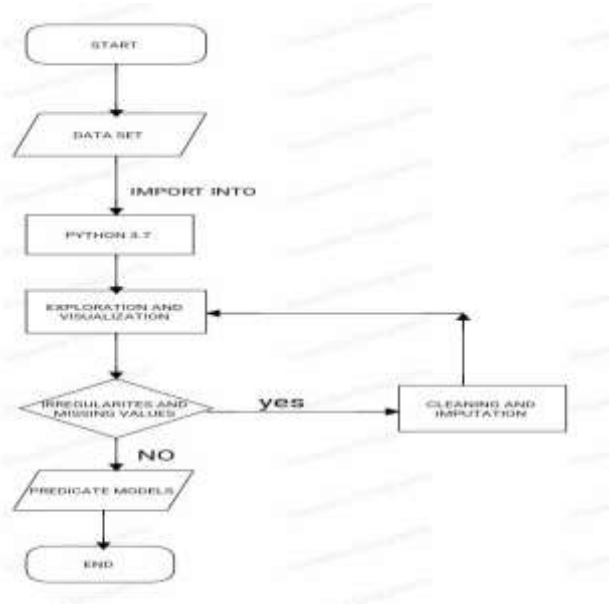


**Figure 1:** Dataset Description of Big Mart

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{x}_{\mathrm{predict}} - \mathrm{x}_{\mathrm{actual}}\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|\mathrm{x}_{\mathrm{predict}} - \mathrm{x}_{\mathrm{actual}}\right|\right)^2}$$

The dataset is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store capacity. hypotheses involves attributes like: brand, advertisement, promotional offer, etc. After considering all, a dataset is formed and finally the data-set was divided into two parts, training set and test set in the ratio 80: 20.

### *3.1 Data Exploration:*

These values may not be appropriate in this form. So, we need to convert them into how old a particular outlet is. There are 1559 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item type contains 16 unique values. Where as two types of Item Fat Content are there but some of them are misspelled as regular instead of 'Regular' and low fat, LF instead of Low Fat. From Figure 2. It was found that the response variable i.e. Item Outlet Sales was positively skewed. So, to remove the skewness of response variable a log operation was performed on Item Outlet Sales.

### *3.2 Data Cleaning:*

It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model we are assuming that there is no relationship between the measured attribute and imputed attribute.

### *3.3 Feature Engineering:*

Some nuances were observed in the data-set during data exploration phase. So this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. In some cases, it was noticed that non-consumables and fat content property are not specified. To avoid this we create a third category of Item fat content i.e. none. In the Item Identifier attribute, it was found that the unique ID starts with either DR or FD or NC. So, we create a new attribute Item Type New with three categories like Foods, Drinks and Non-consumables. Finally, for determining how old a particular outlet is, we add an additional attribute Year to the dataset.

### *3.4 Model Building:*

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is build it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model using Decision tree algorithm and compare it with other machine learning techniques like linear regression, Ridge regression.

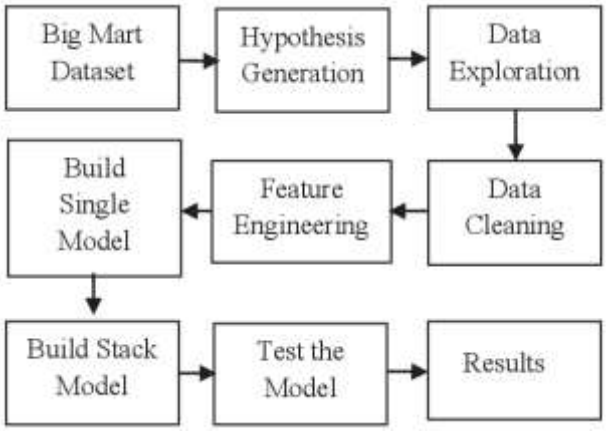


**Figure 2:** predictive model to forecast sales of Big Mart

## 4. Results and Discussion:

### *4.1 Linear Regression:*

A model which create a linear relationship between the dependent variable and one or more independent variable, mathematically linear regression

### *4.2 Decision Tree Regression:*

Decision trees are basically predictive machine learning models. Decision trees models helps to predict a class for the case after training pruning and testing is over.

| Model | Cross Validation Score (Mean) | Cross Validation Score(Std) |
|---|---|---|
| Linear Regression | 1129 | 43.24 |
| Decision Tree | 1091 | 45.42 |
| Ridge Regression | 1097 | 43.41 |

**Table 1:** Comparison of Cross Validation Score (Mean) and Cross Validation Score(Std)

| Model | MAE | RMSE |
|---|---|---|
| Linear Regression | 836.1 | 1127 |
| Decision Tree | 741.6 | 1058 |
| Ridge Regression | 836 | 1129 |

**Table 2:** Comparison of MAE and RMSE of proposed model with other model

## 5. Conclusion:

We have analyzed datasets of big mart sales prediction and performed literature survey related to sales prediction using various techniques .We used Jupyter tool through Anaconda Navigator for processing the techniques. Decision tree based Regression proved the best model to predict the future sales with the accuracy rate. Training the model was easier than any other models. It proved to be the best model in forecasting sales of Big Mart. This indirectly helps to gain more profit and have a scheduled products in stock.

**7. References:**


1. Kommineni, K. K. ., & Prasad, A. . (2023). A Review on Privacy and Security Improvement Mechanisms in MANETs. International Journal of Intelligent Systems and Applications in Engineering, 12(2), 90–99. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/4224
2. Vellela, S.S., Balamanigandan, R. Optimized clustering routing framework to maintain the optimal energy status in the wsn mobile cloud environment. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023- 15926-5
3. Vellela, S. S., Reddy, B. V., Chaitanya, K. K., & Rao, M. V. (2023, January). An Integrated Approach to Improve E-Healthcare System using Dynamic Cloud Computing Platform. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 776-782). IEEE.
4. K. N. Rao, B. R. Gandhi, M. V. Rao, S. Javvadi, S. S. Vellela and S. Khader Basha, "Prediction and Classification of Alzheimer's Disease using Machine Learning Techniques in 3D MR Images," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 85-90, doi: 10.1109/ICSCSS57650.2023.10169550.
5. VenkateswaraRao, M., Vellela, S., Reddy, V., Vullam, N., Sk, K. B., & Roja, D. (2023, March). Credit Investigation and Comprehensive Risk Management System based Big Data Analytics in Commercial Banking. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 2387-2391). IEEE [6]
6. S Phani Praveen, RajeswariNakka, AnuradhaChokka, VenkataNagarajuThatha, SaiSrinivasVellela and UddagiriSirisha, "A Novel Classification Approach for Grape Leaf Disease Detection Based on Different Attention Deep Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 14(6), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01406128
7. Vellela, S. S., & Balamanigandan, R. (2022, December). Design of Hybrid Authentication Protocol for High Secure Applications in Cloud Environments. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 408-414). IEEE.
8. Vullam, N., Vellela, S. S., Reddy, V., Rao, M. V., SK, K. B., & Roja, D. (2023, May). Multi-Agent Personalized Recommendation System in E-Commerce based on User. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1194-1199). IEEE.
9. Vellela, S. S., Balamanigandan, R., & Praveen, S. P. (2022). Strategic Survey on Security and Privacy Methods of Cloud Computing Environment. Journal of Next Generation Technology (ISSN: 2583-021X), 2(1).
10. Vellela, S. S., & Krishna, A. M. (2020). On Board Artificial Intelligence With Service Aggregation for Edge Computing in Industrial Applications. Journal of Critical Reviews, 7(07), 2020.
11. Madhuri, A., Jyothi, V. E., Praveen, S. P., Sindhura, S., Srinivas, V. S., & Kumar, D. L. S. (2022). A New Multi-Level Semi-Supervised Learning Approach for Network Intrusion Detection System Based on the 'GOA'. Journal of Interconnection Networks, 2143047.
12. Madhuri, A., Praveen, S. P., Kumar, D. L. S., Sindhura, S., &Vellela, S. S. (2021). Challenges and issues of data analytics in emerging scenarios for big data, cloud and image mining. Annals of the Romanian Society for Cell Biology, 412-423.
13. Praveen, S. P., Sarala, P., Kumar, T. K. M., Manuri, S. G., Srinivas, V. S., &Swapna, D. (2022, November). An Adaptive Load Balancing Technique for Multi SDN Controllers.In 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 1403-1409).IEEE.
14. Vellela, S. S., Basha Sk, K., & Yakubreddy, K. (2023). Cloud-hosted concept-hierarchy flex-based infringement checking system. International Advanced Research Journal in Science, Engineering and Technology, 10(3).
15. Rao, M. V., Vellela, S. S., Sk, K. B., Venkateswara, R. B., & Roja, D. (2023). SYSTEMATIC REVIEW ON SOFTWARE APPLICATION UNDERDISTRIBUTED DENIAL OF SERVICE ATTACKS FOR GROUP WEBSITES. Dogo Rangsang Research Journal UGC Care Group I Journal, 13(3), 2347-7180.
16. Venkateswara Reddy, B., Vellela, S. S., Sk, K. B., Roja, D., Yakubreddy, K., & Rao, M. V. Conceptual Hierarchies for Efficient Query Results Navigation. International Journal of All Research Education and Scientific Methods (IJARESM), ISSN, 2455-6211.
17. Sk, K. B., Roja, D., Priya, S. S., Dalavi, L., Vellela, S. S., & Reddy, V. (2023, March). Coronary Heart Disease Prediction and Classification using Hybrid Machine Learning Algorithms. In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (pp. 1-7). IEEE.
18. Sk, K. B., & Vellela, S. S. (2019). Diamond Search by Using Block Matching Algorithm. DIAMOND SEARCH BY USING BLOCK MATCHING ALGORITHM. International Journal of Emerging Technologies and Innovative Research (www. jetir. org), ISSN, 2349-5162.
19. Yakubreddy, K., Vellela, S. S., Sk, K. B., Reddy, V., & Roja, D. (2023). Grape CS-ML Database-Informed Methods for Contemporary Vineyard Management. International Research Journal of Modernization in Engineering Technology and Science, 5(03).

20. Vellela, Sai Srinivas and Chaganti, Aswini and Gadde, Srimadhuri and Bachina, Padmapriya and Karre, Rohiwalter, A Novel Approach for Detecting Automated Spammers in Twitter (June 24, 2023). Mukt Shabd Journal Volume XI, Issue VI, JUNE/2022 ISSN NO : 2347-3150, pp. 49-53 , Available at SSRN: https://ssrn.com/abstract=4490635

21. Vellela, Sai Srinivas and Pushpalatha, D and Sarathkumar, G and Kavitha, C.H. and Harshithkumar, D, ADVANCED INTELLIGENCE HEALTH INSURANCE COST PREDICTION USING RANDOM FOREST (March 1, 2023). ZKG International, Volume VIII Issue I MARCH 2023, Available at SSRN: https://ssrn.com/abstract=4473700

22. Dalavai, L., Javvadi, S., Sk, K. B., Vellela, S. S., & Vullam, N. (2023). Computerised Image Processing and Pattern Recognition by Using Machine Algorithms.

23. Vellela, S. S., Basha Sk, K., & Javvadi, S. (2023). MOBILE RFID APPLICATIONS IN LOCATION BASED SERVICES ZONE. MOBILE RFID APPLICATIONS IN LOCATION BASED SERVICES ZONE", International Journal of Emerging Technologies and Innovative Research (www. jetir. org| UGC and issn Approved), ISSN, 2349-5162.

24. Vellela, Sai Srinivas and Sk, Khader Basha and B, Venkateswara Reddy, Cryonics on the Way to Raising the Dead Using Nanotechnology (June 18, 2023). INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS), Vol. 03, Issue 06, June 2023, pp : 253-257,

25. Vellela, Sai Srinivas and D, Roja and B, Venkateswara Reddy and Sk, Khader Basha and Rao, Dr M Venkateswara, A New Computer-Based Brain Fingerprinting Technology (June 18, 2023). International Journal Of Progressive Research In Engineering Management And Science, Vol. 03, Issue 06, June 2023, pp : 247-252 e-ISSN : 2583-1062.,

26. Gajjala, Buchibabu and Mutyala, Venubabu and Vellela, Sai Srinivas and Pratap, V. Krishna, Efficient Key Generation for Multicast Groups Based on Secret Sharing (June 22, 2011). International Journal of Engineering Research and Applications,Vol. 1, Issue 4, pp.1702-1707, ISSN: 2248-9622

27. Kiran Kumar Kommineni, Ratna Babu Pilli, K. Tejaswi, P. Venkata Siva,Attention-based Bayesian inferential imagery captioning maker,Materials Today: Proceedings,2023,ISSN 2214-7853,https://doi.org/10.1016/j.matpr.2023.05.231.

28. Venkateswara Reddy, B., &KhaderBashaSk, R. D. Qos-Aware Video Streaming Based Admission Control And Scheduling For Video Transcoding In Cloud Computing. In International Conference on Automation, Computing and Renewable Systems (ICACRS 2022).

29. Reddy, N. V. R. S., Chitteti, C., Yesupadam, S., Desanamukula, V. S., Vellela, S. S., & Bommagani, N. J. (2023). Enhanced speckle noise reduction in breast cancer ultrasound imagery using a hybrid deep learning model. Ingénierie des Systèmesd'Information, Vol. 28, No. 4.

30. Vellela, S. S., & Balamanigandan, R. (2023). An intelligent sleep-awake energy management system for wireless sensor network. Peer-to-Peer Networking and Applications, 16(6), 2714-2731.

31. Rao, D. M. V., Vellela, S. S., Sk, K. B., &Dalavai, L. (2023). Stematic Review on Software Application Under-distributed Denial of Service Attacks for Group Website. DogoRangsang Research Journal, UGC Care Group I Journal, 13.

32. Priya, S. S., Vellela, S. S., Reddy, V., Javvadi, S., Sk, K. B., & Roja, D. (2023, June). Design And Implementation of An Integrated IOT Blockchain Framework for Drone Communication. In 2023 3rd International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE.

33. Vullam, N., Yakubreddy, K., Vellela, S. S., Sk, K. B., Reddy, V., & Priya, S. S. (2023, June). Prediction And Analysis Using A Hybrid Model For Stock Market. In 2023 3rd International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE.

34. K. K. Kumar, S. G. B. Kumar, S. G. R. Rao and S. S. J. Sydulu, "Safe and high secured ranked keyword searchover an outsourced cloud data," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 2017, pp. 20-25, doi: 10.1109/ICICI.2017.8365348.

35. Sk, K. B., Vellela, S. S., Yakubreddy, K., & Rao, M. V. (2023). Novel and Secure Protocol for Trusted Wireless Ad-hoc Network Creation. Khader Basha Sk, Venkateswara Reddy B, Sai Srinivas Vellela, Kancharakunt Yakub Reddy, M Venkateswara Rao, Novel and Secure Protocol for Trusted Wireless Ad-hoc Network Creation, 10(3).

36. Vellela, S. S., Sk, K. B., Dalavai, L., Javvadi, S., & Rao, D. M. V. (2023). Introducing the Nano Cars Into the Robotics for the Realistic Movements. International Journal of Progressive Research in Engineering Management and Science (IJPREMS) Vol, 3, 235-240.

37. Kumar, K. & Babu, B. & Rekha, Y.. (2015). Leverage your data efficiently: Following new trends of information and data security. International Journal of Applied Engineering Research. 10. 33415-33418.

38. Vellela, S. S., Reddy, V. L., Roja, D., Rao, G. R., Sk, K. B., & Kumar, K. K. (2023, August). A Cloud-Based Smart IoT Platform for Personalized Healthcare Data Gathering and Monitoring System. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5). IEEE.

39. Davuluri, S., Kilaru, S., Boppana, V., Rao, M. V., Rao, K. N., & Vellela, S. S. (2023, September). A Novel Approach to Human Iris Recognition And Verification Framework Using Machine Learning Algorithm. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 6, pp. 2447-2453). IEEE.

40. Vellela, S. S., Vuyyuru, L. R., MalleswaraRaoPurimetla, N., Dalavai, L., & Rao, M. V. (2023, September). A Novel Approach to Optimize Prediction Method for Chronic Kidney Disease with the Help of Machine Learning Algorithm. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 6, pp. 1677-1681). IEEE.

41. Vellela, S. S., Roja, D., Sowjanya, C., SK, K. B., Dalavai, L., & Kumar, K. K. (2023, September). Multi-Class Skin Diseases Classification with Color and Texture Features Using Convolution Neural Network. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 6, pp. 1682-1687). IEEE.

42. Vellela, S. S., Sk, K. B., & Reddy, V. An Intelligent Decision Support System for retrieval of patient's information.

43. Rao, M. V., Sreeraman, Y., Mantena, S. V., Gundu, V., Roja, D., & Vatambeti, R. (2023). Brinjal Crop yield prediction using Shuffled shepherd optimization algorithm based ACNN-OBDLSTM model in Smart Agriculture. Journal of Integrated Science and Technology, 12(1), 710. Retrieved from https://pubs.thesciencein.org/journal/index.php/jist/article/view/a710

44. Vellela, S. S., Narapasetty, S., Somepalli, M., Merikapudi, V., & Pathuri, S. (2022). Fake News Articles Classifying Using Natural Language Processing to Identify in-article Attribution as a Supervised Learning Estimator. Mukt Shabd Journal, 11.