Leveraging Big Data Analytics for Predictive Modeling and Forecasting in Agriculture in Zambia

Calvin Swatulani Silwizya

School of ICT. Zambia University College of Technology, Ndola, 10101, Zambia

csilwizya@zut.edu.zm

Chiyaba Njovu

School of Computing Technology and Applied Sciences.

ZCAS University, Lusaka, 10101, Zambia

chiyaba.njovu@zcasu.edu.zm

Bob Jere

School of Computing Technology and Applied Sciences.

ZCAS University, Lusaka, 10101, Zambia

Abstract - Big Data is defined using characteristics and concepts beyond size, pinpointing to the volume, velocity, variety, and veracity of the data. The integration of big data analytics in agriculture is revolutionizing farming practices, crop management, and decision-making processes. Much of the existing research has utilized limited datasets and simplistic analytical methods, such as basic statistical approaches and opaque machine learning models, which hinder clear interpretation by farmers and stakeholders. The study aimed to develop a predictive model and forecasting accuracy using data analytics that will improve crop yield in Agriculture, applied advanced data analytics approaches with tree-based machine learning techniques to pinpoint key factors that influence agricultural productivity and used key factors to build a model that predicts crop yield. The study implemented experimental methodology. Utilizing the LightGBM framework - a gradient boosting model known for its interpretability, analyzed an amalgamation of data from surveys, farm records, and climatic information to assess feature importance. It also integrated diverse datasets from governmental reports and agricultural archives. This analysis included various socio-economic factors such as access to water, soil quality, type of seeds, weather pattern, educational levels of farmers, and market access, which were identified as critical variables affecting agricultural success. The LightGBM model not only achieved high accuracy and reliability but also provide transparent insights, outperforming other methods like XGBoost, decision trees, and random forests in our evaluations.

Keywords: Big data, LightGBM, Model, Decision tree, Data analytics

INTRODUCTION

Big Data analytics in the agricultural sector has huge potential to contribute to the requirements of food production. Predictive analytics is a term that covers in principle, the same area as predictive modelling, but in practice it is also used to describe general trends in advanced data processing. The study demonstrates the role of Big Data in pertinent data acquisition from factors affecting the agriculture sector, such as climatic and weather, soil and land, crop variety, agronomic practices, pests, diseases and biological, social, economic and management factors and technological factors. In this study crop surveys and climate data from 2001 to 2024 were be collected, analysed, cleaned and integrated, a dataset was built used for predictive analysis and forecasting.

Predictive analytics emphasizes on building models that result in fit statistics. This was used to perceive how crop growth is sensitive to climate factors, soil conditions, and farming practices [1]. Using precision agriculture through predictive models is the concept that enables farmers to understand crops at the micro level and manage the crops smartly [2]. Therefore, Big data analytics was implemented and effectively used it to develop predictive model tailored to analyse and improve agricultural landscape. The Agricultural sector faces numerous challenges, including unpredictable weather patterns, low productivity, poor resource allocation, and fluctuating market prices. These factors hinder the ability of farmers to plan effectively and reduce production losses, and ensure food security. By harnessing big data, the agricultural sector can gain insights that support better crop management, resource allocation, and risk assessment.

 ISBN: 978-9982-95-500-3

RELATED WORKS

This paper reviews literature on the effectiveness of existing big data analytical models in predicting maximum crop yield, investigates papers on challenges and opportunities of applying predictive models to maximise crop yield, resource allocation, and market prediction, and how multiple data sources can be integrated in developing a big data-driven predictive model.

The study, [3] presented a model on crop yield prediction using machine learning techniques, and extracted major machine learning algorithms, features and evaluation metrics used in the yield estimation by integrating agrarian factors in machine learning techniques [4]. This allowed them to show a strong relationship between crop yield and climatic factors. According to [5] use of computer vision and AI to enhance the grain quality of five crops (maize, rice, wheat, soybean and barley), disease detection and phenotyping. [6] reviewed the application of big data analysis in some fields of agriculture. It highlighted solutions to some key well-known problems, used tools and algorithms, along with input datasets. The authors concluded that big data analytics in agriculture is still at its early stage, and many barriers need to be overcome, despite the availability of the data and tools to analyse it.

Researchers have employed various modeling approaches, including crop simulation models, statistical analysis, agro-economic simulation, and computable general equilibrium models, to quantify the economic impacts of climate change on agriculture globally or in specific regions. [7] These studies have reported substantial differences in outcomes, such as production, trade, welfare, and prices, due to differences in model parameterization and specification [7].

Most crop models found in pre-precision agriculture literature and during its dawn typically are based on linear regression analysis, calculations of root mean square error, and mean error [8]. Multiple linear regression techniques using interaction terms are considered an improvement over strictly linear models. Multiple linear regression and linear mixed models are used in soil mapping, where the variability of a target soil property is explained by its relationships with other soil and climate factors, with shortcomings like autocorrelation and non-linearity between variables. In their paper [9] stated that the high complexity and non-linearity of problems faced in agriculture require methods able to approximate complex mappings by integrating data coming from different sources and exploiting the information contained in the reference samples. According to [10] having used the Naıve Bayes classifier to learn models and to make predictions. The authors showed that Naıve Bayes has good performance on sparse datasets, extremely fast to run on a large, sparse dataset when it is formulated well. The main speedup stems from the fact that Naıve Bayes completely ignores inter feature dependencies by assuming that within-class covariance's of the features are zero. In the study [28] the author proved the application of supervised machine learning algorithms (Logistic regression and Light gradient booting) to combine data sources in predictive analysis,

The development of big data-driven predictive models is a rapidly evolving field. The studies reviewed shows gaps in the techniques used. Authors [11]; [8] illustrates the use of regression analysis well in predictive analysis but creates a limit only to development of models using linear regression analysis, calculations of root mean square error, and mean error. This technique or method causes challenges when the data contains outliers. In [13] the authors present the use of MapReduce Model to make analysis of different type of parameters that give us what if analysis. In the literature [14] a large gap between potential and actual yield were founded by WOFOST model. The WOFOST model often overestimates, underestimates or slightly offsets the normal estimates.

METHODOLOGY

In this research study an experimental research methodology was used as it allowed the researcher to analyse data, validate the findings and explain unexpected results. The study also adopted a predictive approach. The integration of advanced analytics enabled transformation of data into actionable insights, enabling farmers to anticipate market trends and adapt to environmental changes. For instance, the exploitation of large datasets on local agriculture practices, including climate patterns and soil conditions was employed to significantly improve decision-making processes. However, as highlighted in contemporary analyses, the challenges associated with underutilized data resources persist [15]

 ISBN: 978-9982-95-500-3

The vast data was collected from three (3) provinces namely Luapula, Copperbelt and Southern and then build datasets. This include climate data, soil health information and seeds type that were planted in each farming season with their respective crop yield per area size to anticipate future agricultural outcomes. The data collected in this study is from 2001 to 2024.

The study also gather data relating to the models used in predicting crop yield and further examine the challenges and opportunities of applying predictive models to improve crop yield, resource allocation, and market prediction. The data was cleaned, analysed and a datasets built. The following methodological framework was used as per diagram in Figure 3.1.

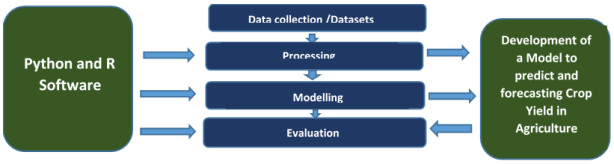

Figure 3.1: Methodological Framework Diagram

The data was collected from various sources, such as government agencies, research institutions, and Metrological Department, to create a comprehensive dataset.

Particular attention given to the quality, completeness, and timeliness of the data, as these factors were critical for effective predictive modeling and forecasting.

MODEL/FRAMEWORK

Accurate predictions of crop yield are critical for effective crop management, resource allocation, and strategic planning in agriculture [16]. The escalating volatility of food prices has underscored the urgency of enhancing crop productivity on existing farmlands to meet the demands of a growing global population, further emphasizing the importance of reliable crop yield models [17].

The proposed model used the following factors Soil type to ascertain the ratios of Nitrogen, Phosphorus, Potassium, Temperature, Humidity and pH values of soil, Crop spacing, seed type, Climate (rainfall pattern), planting periods, crop diseases and market demand. These factor were used by in the model to predict maximum crop yield. This empirical statistical model, would establishing complex relationships between crop yield and related variables.

In the conceptual design the performance of LightGBM was compared with the prediction performance of benchmark models trained using XGBoost, random forest and decision trees.

SHAP was used to interpret the proposed model such that it can be easily understood and validated by end users. The focus was to expand the scope of EDM and also provide actionable insights and models to improve crop yield in Agriculture.

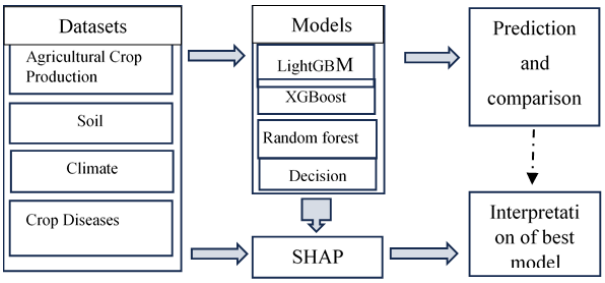

Figure 4.1: Proposed Model

Using SHAP, enabled the interpretation and visualization of the contribution of features. Features on the right tend to push the model prediction to the base value while those on the left pushes the prediction to the output value. In this case, historical agricultural production provides the biggest impact. Therefore, end users of this model such as farmers and government are able to interpret and understand the impact factor of all the control features. To understand the effect of one feature in the prediction, a SHAP value of that feature was be plotted against other feature SHAP values in the dataset as shown in Figure 4.2.1.

CONCLUSION

The integration of machine learning algorithms with Big data analysis technologies in the proposed model offers a paradigm shift in yield forecasting, enabling a transition from conventional, often subjective, methods to more data-driven and objective strategies that leverage complex relationships between environmental factors and crop performance. This model will provide solutions to challenges affection the maximizing and prediction of crop yield which is much needed as a solution to farmer's challenges.

REFERENCES

[1] D. Amanullah and S. Khalid, *Agronomy: Climate Change*. BoD – Books on Demand, 2020.

 ISBN: 978-9982-95-500-3

[2] E. E. K. Senoo *et al.*, "IoT Solutions with Artificial Intelligence Technologies for Precision Agriculture: Definitions, Applications, Challenges, and Opportunities," *Electronics*, vol. 13, no. 10, p. 1894, May 2024, doi: 10.3390/electronics13101894.

[3] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[4] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Comput. Electron. Agric.*, vol. 155, pp. 257–282, Dec. 2018, doi: 10.1016/j.compag.2018.10.024.

[5] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Comput. Electron. Agric.*, vol. 153, pp. 69–81, Oct. 2018, doi: 10.1016/j.compag.2018.08.001.

[6] E. Kamir, F. Waldner, and Z. Hochman, "Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 160, pp. 124–135, Feb. 2020, doi: 10.1016/j.isprsjprs.2019.11.008.

[7] J. Delincé, P. Ciaian, and H.-P. Witzke, "Economic impacts of climate change on agriculture: the AgMIP approach," *J. Appl. Remote Sens.*, vol. 9, no. 1, p. 097099, Jan. 2015, doi: 10.1117/1.JRS.9.097099.

[8] G. Morota, R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando, "big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1," *J. Anim. Sci.*, vol. 96, no. 4, pp. 1540–1550, Apr. 2018, doi: 10.1093/jas/sky014.

[9] K. H. Coble, A. K. Mishra, S. Ferrell, and T. Griffin, "Big Data in Agriculture: A Challenge for the Future," *Appl. Econ. Perspect. Policy*, vol. 40, no. 1, pp. 79–96, 2018, doi: 10.1093/aepp/ppx056.

[10] E. Junqué de Fortuny, D. Martens, and F. Provost, "Predictive Modeling With Big Data: Is Bigger Really Better?," *Big Data*, vol. 1, no. 4, pp. 215–226, Dec. 2018, doi: 10.1089/big.2013.0037.

[11] J. Verrelst *et al.*, "Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties – A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 273–290, Oct. 2015, doi: 10.1016/j.isprsjprs.2015.05.005.

[12] "'Adopting Information System Technologies In Construction Project Manag' by Hadi Haikal." Accessed: Apr. 24, 2025. [Online]. Available: https://digitalcommons.harrisburgu.edu/dandt/5/

[13] K. Bronson and I. Knezevic, "Big Data in food and agriculture," *Big Data Soc.*, vol. 3, no. 1, p. 2053951716648174, Jun. 2016, doi: 10.1177/2053951716648174.

[14] H. L. Boogaard and I. Supit, "System description of the WOFOST 7.2, cropping systems model," 2020.

[15] A. A. Tapo, A. Traore, S. Danioko, and H. Tembine, "Machine Intelligence in Africa: a survey," Feb. 03, 2024, *arXiv*: arXiv:2402.02218. doi: 10.48550/arXiv.2402.02218.

[16] H. T. Pham, J. Awange, M. Kuhn, B. V. Nguyen, and L. K. Bui, "Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices," *Sensors*, vol. 22, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/s22030719.

[17] R. L. de F. Cunha, B. Silva, and P. B. Avegliano, "A Comprehensive Modeling Approach for Crop Yield Forecasts using AI-based Methods and Crop Simulation Models," Jun. 16, 2023, *arXiv*: arXiv:2306.10121. doi: 10.48550/arXiv.2306.10121.

[18] K. Meghraoui, I. Sebari, J. Pilz, K. Ait El Kadi, and S. Bensiali, "Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges," *Technologies*, vol. 12, no. 4, Art. no. 4, Apr. 2024, doi: 10.3390/technologies12040043.

[19] J. Fu *et al.*, "Regionally variable responses of maize and soybean yield to rainfall events in China," *Agric. For. Meteorol.*, vol. 364, p. 110458, Apr. 2025, doi: 10.1016/j.agrformet.2025.110458.

[20] Í. M. da Cunha, R. B. dos Reis, R. B. Silveira, R. O. de Sousa, C. L. R. de Lima, and F. S. Carlos, "Soil Breaking Mechanisms Increase Phosphorus, Potassium, and Water Use Efficiency of Soybeans Under Dry Season in Paddy Fields in Southern Brazil," *Int. J. Plant Prod.*, Mar. 2025, doi: 10.1007/s42106-025-00332-8.

[21] R. de S. Nóia Júnior, L. Olivier, D. Wallach, E. Mullens, C. W. Fraisse, and S. Asseng, "A simple procedure for a national wheat yield forecast," *Eur. J. Agron.*, vol. 148, p. 126868, Aug. 2023, doi: 10.1016/j.eja.2023.126868.

[22] Z. Su *et al.*, "Climate-adaptive crop distribution can feed food demand, improve water scarcity, and reduce greenhouse gas emissions," *Sci. Total Environ.*, vol. 944, p. 173819, Sep. 2024, doi: 10.1016/j.scitotenv.2024.173819.

[23] F. Nobre Cunha, G. Nobre Cunha, M. Batista Teixeira, N. Furtado Da Silva, and A. Antunes Lopes, "PRODUCTION POTENTIAL AND CROP EVAPOTRANSPIRATION ESTIMATION FOR BEAN, SOYBEAN, AND MAIZE USING THE SEBAL ALGORITHM. | EBSCOhost." Accessed: May 01, 2025. [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.15809%2Firriga.2023v28n3p496-506?sid=ebsco:plink:crawler&id=ebsco:doi:10.15809%2Firriga.2023v28n3p496-506

[24] W. Veerakachen and M. Raksapatcharawong, "RiceSAP: An Efficient Satellite-Based AquaCrop Platform for Rice Crop Monitoring and Yield Prediction on a Farm- to Regional-Scale," *Agronomy*, vol. 10, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/agronomy10060858.

 ISBN: 978-9982-95-500-3

[25] Y. Su, Y. Liu, L. Huo, and G. Yang, "Research on optimal allocation of soil and water resources based on water–energy–food–carbon nexus," *J. Clean. Prod.*, vol. 450, p. 141869, Apr. 2024, doi: 10.1016/j.jclepro.2024.141869.

[26] L. Bastiaans, M. J. Kropff, J. Goudriaan, and H. H. van Laar, "Design of weed management systems with a reduced reliance on herbicides poses new challenges and prerequisites for modeling crop–weed interactions," *Field Crops Res.*, vol. 67, no. 2, pp. 161–179, Jul. 2000, doi: 10.1016/S0378-4290(00)00091-5.

[27] R. Sumathi and N. R. Raajan, "A multilevel distributed image based encryption for full integrity," *Multimed. Tools Appl.*, vol. 79, no. 3, pp. 2161–2183, Jan. 2020, doi: 10.1007/s11042-019-08104-z.

[28] H. Wandera, V. Marivate, and M. D. Sengeh, "Predicting National School Performance for Policy Making in South Africa," in *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Nov. 2019, pp. 23–28. doi: 10.1109/ISCMI47871.2019.9004323.

 ISBN: 978-9982-95-500-3