Journal of Applied Sciences, Information and Computing

Volume 4, Issue 1, July 2023

School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2023-4-1-03

DEVELOPMENT OF A MACHINE LEARNING REGRESSION MODEL FOR ACCURATE SUGARCANE CROP YIELD PREDICTION, JINJA – UGANDA

Yuma Erick[1], Chinecheremu Umezuruike[2], Nasasira Jossy[3], Balyejusa Gusite[4]

[1]Directorate of Research, Innovation, Constance and Extensions, Kampala International University, Uganda. ericksyuma@gmail.com
[2]Assistant Professor Software Engineering, Bowen University, umezuruike.Chinecheremu@gmail.com
[3]Department Of Information Technology, Kampala International University, Uganda.
josirah23@gmail.com
[4]Department Of Information Systems, Kampala International University, Uganda.
augsoft.systemz@gmail.com

Abstract

Sugarcane is one of the key crops grown worldwide and used for sugar processing, food, alcohol, biogas, fertilizer, and other products. There is a problem with Sugarcane yield prediction, yields aren't accurately predicted, and this creates an impact on yields. This research looks at identifying methods used for the prediction, design, development, and evaluation of the three machine-learning regression models used for predicting sugarcane yields in Uganda. This research was implemented using Data Science methodology, several machine learning algorithms for prediction of yields on dataset have been analyzed. The collected and analyzed dataset in this research had one output/ dependent variable and eight independent variables. The algorithms used to develop the prediction models are the Multiple Linear Regression algorithm, Decision Tree Regression algorithm, and Random Forest Regression algorithm to predict the output. The dataset of 3 years, 2019, 2020, and 2022 was considered and merged to train and test the model at a ratio of 80% to 20%. The accuracies of the individual models were compared after training, testing the dataset, and evaluation. The multiple Linear regression model results indicate that out of 100%, the model accuracy was 76.5%, the Decision Tree Regression Model scored 89.2%, Random Forest Regression Model was 94.6%. The random forest model came out as the best model. The Random Forest model has a percentage improvement of 60.4%. In future research, researchers can work

on, A web-based machine learning model, Deep learning methods used to improve the model and more data can be used to improve the accuracy.

Keywords: Machine Learning, Kakira Sugar Limited, Random Forest Regression, Decision Tree Regression, Machine Learning Algorithm, Multiple Linear Regression

1. INTRODUCTION

Agricultural Organization in UN stated that 90 and above countries grow sugarcanes in an area of 26 million hectares and 1.8 billion tons of global harvests of sugarcane (Taherei Ghazvinei et al., 2018). Brazil is the leading sugar producer in the world with 42 million metric tons in 2020/2021 mounting to 179.66 million metric tons worldwide, followed by India, China, Cuba, Pakistan, Mexico, Iran, and Thailand (Vijayakumar & Bozward, 2021, Shahbandeh, 2021).

Uganda is the leading sugarcane grower in East Africa, followed by Kenya and Tanzania. The agricultural sector is one of the vital parts of the economy in Uganda because millions of people are engaged in it. Sugarcane is one of the main crops grown in Uganda and is a significant contributor to the country's economy. In 2020, sugarcane production for Uganda was 5.78 million tonnes growing at an average annual rate of 3.76%.

1.1 Sugarcane Production

Sugarcane farmers have multiple fields with canes at different stages of growth to make sure there is always availability of sugarcane to supply the mill. The cane is planted and all through to harvest, the process can take from 8 months to 18 months in the fields. This makes sugarcane a slow perennial crop to produce. Preparation of sugarcanes before planting is very vital. The cane leaves are removed and therefore cut into 20cm pieces. The field is prepared with incisions made at approximately 15cm to 20cm in width. The pieces of sugarcane are placed in the horizontal holes, then covered with soil and left to grow. The production of cane starts with healthy soil. This is done by adding nutrients and water. It can take 12 to 18 months approximately to mature. During this time, the cane must be treated with pesticides and fertilizers to ensure healthy and high-yield sugarcane.

The problem faced by most farmers growing sugarcane is that they don't follow the weather patterns and soil structure before choosing where to plant the crop. They don't invest time and money to research the conditions favorable to bring out good yields in the area they want to produce the crop.

1.2 Sugarcane harvest

Sugarcane at times doesn't need replanting because it's the top of the plant which is removed while harvesting. It is harvested either manually or mechanized. Manual harvest methods can easily make the sugarcane gardens to be burnt. All the leaves are removed so that sugarcane is then manually cut to the ground then removed and taken to the mill by tractors or other means of transport. The mechanized harvest uses machines to extract the sugarcane as it moves in the garden, while it is loaded into a truck. The mechanized harvesting method is seen as the future of cane harvesting. (Spencer, February 21, 2020).

Sugarcane also called Saccharum officinarum is a perennial species of tall grass of Asian native whose growth has lasted for more than 4000 years planted for the juice where sugar processing is done and doesn't need to be re-planted every year. When harvesting, the stalk is cut above the root level so that it can regrow. In estimation, the plant crop is expected to yield 100%, Ratoon 1 – 70%, Ratoon 2 – 60%, and Ratoon 3 – 50% tons for Kakira Sugar Limited. The plant height reaches 3–6 meters, it is also used for direct consumption as food. The world's most sugar cane is cultivated in subtropical and tropical places. Biofuel production, alcohol,

https://doi.org/10.59568/JASIC-2023-4-1-03

fertilizer, and electricity generation are also obtained from sugarcane.

Sugarcane manufacturing methods have been designed in India by 400 BC. By the 11th century AD, sugar was introduced to Europeans during the Crusades and then taken all over the European countries, the gentleman called Christopher Columbus is said to have imported sugarcane to the West Indies and more than 75% of world's sugar is got from sugarcane today (Sugarcane Profile, 2021). Sugarcane production in Uganda started in the 1920s in the Busoga region introduced by Muljibhai Madhivan with 320 hac in Kakira sub-county in Jinja. sugarcane growth regions in Busoga (KSL) with 10,000 hectares of nucleus estate 856 hectares of satellite estates and over 4,400 hectares at Kayunga estate.

The main manufacturers of sugar in Uganda are Kakira Sugar Works, Kinyara Sugar Works Limited, GM Sugar Uganda Limited, Sango Bay Estates Limited, Sugar Corporation of Uganda Limited and others include the following among others: Amuru Sugar Works Limited, Atiak Sugar Factory, Bugiri Sugar Factory, Busia Sugar Limited, Hoima Sugar Limited, Kamuli Sugar Limited, Kenlon Industries Uganda Limited, Mayuge Sugar Industries Limited, Mukwano Sugar Factory, Seven Star Sugar Limited, Sezibwa Sugar Limited, Buikwe Sugar Works Limited, Sugar & Allied Industries Limited, Uganda Farmers Crop Industries Limited.

1.3 Problem Statement

In Uganda, there is a problem with Sugarcane crop yield prediction. The yields are not accurately predicted and this creates an impact on sugarcane yield management.

At Kakira Sugar Limited Agronomy Section, there is a problem with sugarcane crop yield prediction. They have challenges related to sugarcane growing and sugar production due to that. There are reasons for the problems of yield fluctuation in the last decade and there are factors responsible. Numerous factors are in play and affect sugarcane crop yields in Uganda and this has an impact on the final yields on maturity. They include; (i) weather for example; Rainfall, temperature, humidity, floods, and others. (ii) soil structure for example; loam soil and sandy soil. (iii) sugarcane variety. (iv) Distance of plantations. (v) planting area. (vi) Age of sugarcane. (vii) irrigation and drainage system. (viii) Pest and diseases, (ix) Fertilizers and fires. (x) method of harvesting and others. (Herbet, 2020)

These cause a problem of fluctuation and inaccurate yields of sugarcanes which brings about low sugar prices, changing production trends of sugar, high cost of production, high prices payable to farmers, high cost of Production, use of unreliable and old machinery, mill sizes, fluctuating prices of canes, inaccurate weeding and harvesting timing, financial crisis. The problem brought by the inaccurate prediction methods used cause difficulty in getting the accurate results of sugarcane yield. The failure to get an accurate prediction of sugarcane yield leads to challenges in planning, decision-making, and budgeting for resources among the sugarcane farmers, factories, and organizations.

1.4 Justifications of the Study

ML models have successfully been implemented in the prediction of crop yields around the world, the algorithms modeled include; MLR, RF, Decision Tree, K Nearest Neighbor, neural networks, Logistic regression, support Vector Machine and K-means, and Convolutional Neural Networks, Recurrent Neural Networks and interaction-based models among others.

Sugarcane crop yield prediction is highly significant since it provides insights that guide companies, industries, and organizations in improving informative knowledge to the stakeholders for mill management, sugarcane cultivation monitoring, sugar production monitoring, informed economic and management decisions, planting and crushing season, corruption, availability and balance of fertilizer.

Research shows the relevance of machine learning and indicates that it provides predictions that are reasonable with higher flexibility and faster results. several aspects are to be investigated concerning accuracy prediction, time taken to predict, supply

https://doi.org/10.59568/JASIC-2023-4-1-03

management, human capital, and machinery management among others. These organizations and companies have engaged a manual approach to yield prediction, which has not proven to be effective and accurate. Hence, there arises a need for accurate sugarcane yield prediction using Machine learning.

1.5 Related Literatures.

Dimo Dimova looked at sugarcane yield estimation through remote sensing time series and phenology metrics. The existing regression-based crop yield estimation approaches used before relying on sets of predictor variables that are specific but have not been compared well. The research compares and demonstrates the use of three sets of object-based predictors for sugarcane yields in the platform of agricultural monitoring knowledge. This uses earth observation data of sentinel-2 satellites captured between 2018 and 2019 for an area of 10,000 hectares in Ethiopia. Many regression models were used. (Dimo Dimova, 2022).

Accurate crop prediction was done using the K-Nearest Neighbor algorithm in Mangalore, Kodagu, Kasarago, and other districts of Karnataka state. The real-time environmental parameters like soil type, Rainfall, humidity, irrigation type, previous yields, location, price, year, type of crop, crop diseases, and its symptoms datasets in India were collected. Related Data on commonly grown crops in the region like coconut, Cardamom, coffee, Areca nut, Ginger, Tea, Paddy, Ground nut, Black gram, Cashew, and Pepper was collected. (Karthikeya et al., 2020).

In soft computing, support vector machines have acquired reasonable significance in making predictions. In this paper, SVM models based on classification were developed to predict rice yields in India. The historical Data on rice production for the years 1950 to 2014 was obtained from the Directorate of Economics and Statistics, Ministry of Agriculture, Indian government. The prediction which was accurate for the four-year relative average increase was achieved as 75.06% using a 4fold cross-validation Method. The experimental software used for this work was MATLAB. This work can be improved by redefining training patterns and learning parameters can be optimized using partial swarm and other techniques to improve its accuracy (Sunil Kumar, 2019).

Forecasting corn yield with ML ensembles was done in three US corn Belt states Indiana, Illinois, and Iowa with datasets that comprise environmental data like weather, soil, and features of management of the 2 distinct scenarios; season partial knowledge, in-season weather complete knowledge and 3 scales; agricultural district, county and state level. Historical yields of corn at the county level from USDA National Agricultural Statistics were obtained from 2000 to 2018. Methods used include; LR, Bayesian search, LASSO Regression, XGBoost and LightGBM, Random Forest, and stacked Generalization. After prediction was made for the ensembles, RL and LASSO regression are the models whose prediction overestimated true values, and others underestimate corn yields. (Shahhosseini et al., 2020).

The study was done by Mupangwa et al., 2020 to answer the following questions. (i) Can ML techniques predict maize grain yields under conservation agriculture? (ii) How close can ML algorithms predict maize grain yields under conservation agriculture-based cropping systems in highlands and lowlands of Eastern Africa and Southern Africa. The maize yield dataset for seven or more years was collected from many websites and 80% was used in training and 20% in testing the algorithm (Mupangwa et al., 2020).

Random Forest Algorithm (RF) has been used to predict accurate sugarcane yield at Tully, in Northeastern Australia. Annual variation in productivity of sugarcane and suitability of predictor variables generated from crop models coupled with observed climate and climate prediction indices (Everingham et al., 2016).

(Wickramasinghe et al., 2021). Looks at modeling the relationship between rice yield and climate variables using statistical and ML techniques. Several climate variables were put under consideration for the application of both methods. Regression techniques, Support Vector Machines, and Artificial Neural Networks are the techniques applied in this relationship modeling. Data collected included rice

https://doi.org/10.59568/JASIC-2023-4-1-03

harvest, yield, and climate data for 2 districts in Sri Lanka for more than 3 decades.

2. METHODOLOGY

2.1 Data Science

Data science is a field of study that uses scientific methods, processes, systems, and algorithms to extract insights and knowledge from structured and unstructured data. It looks at varieties of data by using modern tools and techniques to find patterns and derive knowledge and business decisions. There are more fields under data science, these include; Artificial Intelligence, Deep Learning, and Machine Learning. (Logallo, 2019)

Data science methodology has a routine of finding solutions to specific problems using a cyclic process to guide data scientists which has the following Steps: Business Understanding, Analytic Understanding, Data Requirements, Data Collection, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment, Feedback

Analysis was performed to eliminate noise and outliers from the data. The necessary variables to enable prediction were selected by use of feature selection methods which included the filter method which drops features according to the correlation with the output variable and the wrapper method which split the dataset and trained the model by adding and subtracting the features.

2.2 Methods Used for Sugarcane Yield Prediction in Uganda

Observation Method.

In Kakira Sugar Limited, the fields have attached supervisors who monitor and take care of the sugarcanes. Since they are in touch with the farmers and can observe how the sugarcanes grow for specific gardens. With their observation by looking at the entire field of the sugarcane, they can tell how many trucks of sugarcane can be got. They consider the size of the stems of the sugarcanes and the area covered. The observation method gives an accuracy of 28% of the yields.

Statistical Calculation Using Excel.

In the statistical method, the area and cycle of canes are the major factors of yields that are considered. The field area is the space the sugarcane plant covers and the cycle of the canes is the number of times the plant is left to germinate from the stacks after germinating.

The area is in hectors and the cycles are called ratoons, the first ratoon(R1) gives 100%, R2 gives 70%, and R3 gives 60%.

The prediction is made by multiplying the area of the canes and the ratoon of the cycle. The accuracy gives got is 59% of the sugarcane yields.

That is:

*Prediction = Area * Ratoon of The Crop.*

2.3 Design a Predictive Model for Sugarcane Yield

The design of the predictive model consists of the steps undertaken to make the model meet the main objective of predicting the sugarcane crop yields as shown in the conceptual diagram (Framework) below.

Machine Learning Prediction Model Design

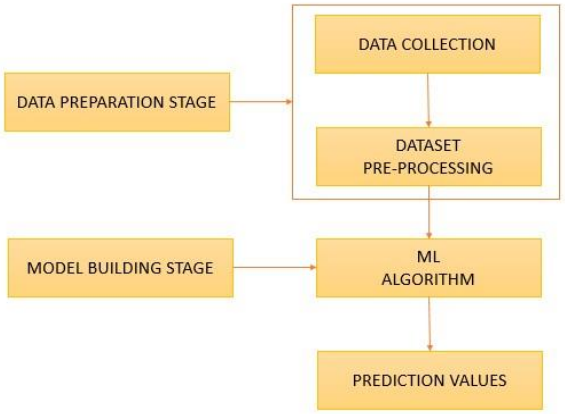

Figure 1 Machine Learning Prediction Model Design

https://doi.org/10.59568/JASIC-2023-4-1-03

Data Collection

The main variables that affect sugarcane crop yields at Kakira Sugar Limited in Jinja district are environment, weather, and management which comprise different components of the mentioned variables like soil structure, rainfall, temperature, area, and others. Since some of the gardens for sugarcane in some locations are not irrigated, irrigation was not taken as a feature to keep the consistency of the dataset. The sugarcane crop yield data for three years were obtained from KSL databases and weather data from the weather station at the industry in Kakira sub-county located in Jinja. The variables are explained below.

1. Actual Sugarcane yields for the previous 3 years after harvest.
2. Area of plant cane per plot.
3. Age of the sugarcane.
4. Sugarcane crop cycle (PC = 100, R1=80, R2=60, R3=40)
5. Temperature in degrees Celsius.
6. Rainfall in millimeters.
7. Soil Potassium.
8. Soil PH.
9. Soil silt.

Data Preprocessing

The following pre-processing operations were considered on the created dataset to make it ready to train the proposed ML model.

- Removal of weather features which are before planting and after harvesting (out of season).
- Aggregation of the weather dataset.
- Observations with lower yields were considered outliers and were left out of the dataset.
- The independent variables were scaled to fit in the set range.
- Feature selection from the dataset.

Dataset Splitting

The dataset used in the ML model was split into 2 portions, train set and test set. This was done in the ratio of 80% Training and 20% testing sets.

Model Training

After preprocessing the collected data, it is split into a train set and a test set. The training set is fed into the developed algorithm. The data was processed by the Algorithm and the model is output which can get the target value in the data provided. The model was trained to develop a learned model. The prediction is made by the model by providing the other attributes of the dataset.

Model Evaluation

The model developed is tested to find out if it can predict the target value quickly and accurately. This can be achieved through model tuning done by optimization of the model parameters to get the Algorithm's best performance. R^2 Score was used to get the accuracy.

Flow Chart for Sugarcane Prediction Model

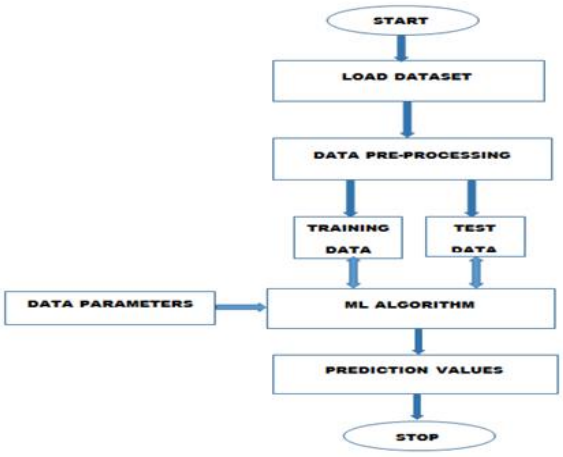

Figure 2 Flow Chart for Sugarcane Prediction Model

The flow chart above shows how the model operations are undertaken when predicting the values of the sugarcane yields from the start to the end.

The model starts by importing the necessary libraries and loading the dataset, preprocessing the data by choosing the most important variables that predict the yields of sugarcane, training the model using 80% of the dataset, training the model using 20% of the dataset, and then predicting the model.

https://doi.org/10.59568/JASIC-2023-4-1-03

2.4 Exploratory Data Analysis

EDA is an approach used to analyze the datasets to summarize the main characteristics and patterns with visual and non-visual methods. It was applied to know and understand the data, it helped the analyst to spend less time on coding and focus more on the data. After data collection, it underwent processing before EDA was done. The cleaned dataset and knowledge from EDA were used to perform modeling and reporting.

Exploratory Data Analysis is as follows;

- Import the necessary required tools and libraries in data analysis.

Pandas is a Python library used to examine, manipulate, and clean the tabular data.

Numpy library which offered special arrays which are different in storage and handling. It requires less space and its first in terms of speed.

Matplotlib library which was used to visualize our dataset to know the patterns.

The script-learn library is a collection of advanced machine-learning algorithms used by Python.

Seaborn was used to visualize data based on matplotlib and also provides a high-level interface for drawing informative statistical and attractive

data = pd.read_csv ("tryy.csv") was used to assign the dataset to the variable data which was used to train and test the model.

- Dataset preview

```
In [21]: data
```

Out[21]:

	Area(Ha)	Crop Cycle	Age	RainFall	Temparature	Soil PH	Soil Silt	k f	Yield
0	14.96	100	34.00	82.5	28.8	7.630	25.410	323.010	573.0
1	14.80	100	26.00	86.1	24.7	7.848	35.407	322.007	555.0
2	14.90	100	13.00	85.8	24.4	7.726	35.404	323.004	550.0
3	14.90	100	23.00	85.8	24.4	5.824	31.401	320.001	550.0
4	14.90	100	24.00	81.7	28.0	6.822	35.398	322.998	550.0
...	...	...	...	...	...	...	...	...	...
840	0.15	100	11.00	81.7	28.0	5.050	20.950	165.950	22.0
841	0.14	70	11.04	81.7	28.0	4.480	22.900	155.900	21.0
842	14.30	30	21.00	81.7	28.0	4.660	24.850	179.850	19.5
843	1.96	45	14.10	85.8	24.4	4.440	22.800	165.800	18.4
844	13.70	30	17.00	85.8	24.4	4.420	27.750	185.750	10.5

Figure 3 Dataset preview

- Check the number of entries, data types, and column types using built-in functions.

The dataset has 844 records which were selected after removing outliers out of 1400 records of data.

- Get the mean, standard deviation, minimum values, and maximum values.
- Get insights of the value numbers in each column which give information about null and duplicate data.
- Plot graphs to get information about the variables.

2. Results

Accuracy of the Machine Learning Models Used

No	Model	Accuracy
1	Multiple Linear Regression	76.5%
2	Decision Tree Regression	89.2%
3	Random Forest Regression	94.6%

Figure 4 Accuracy of the Machine Learning Models Used

https://doi.org/10.59568/JASIC-2023-4-1-03

Random Forest Machine Learning Model Predicted Results

No	Area(Ha)	Crop Cycle	Age	Rainfall	Temperature	Soil PH	Soil Silt	k f	Actual Yields	Predicted Yields
1.	14.1	100	31	86.6	25.2	6.8	30.371	316.971	530	526.4
2.	17.9	100	9.12	84.2	20.1	6.08	34.427	325.827	347.3333	381.8
3.	9.3	70	24	81.7	28	6.68	34.385	258.785	330.5	307.5
4.	6	100	9	81.5	27.8	6.51	30.86	264.56	300	297.1
5.	11.37	50	17.7	84.4	20.3	6.35	30.49	232.29	257.4	261.8
6.	4.07	100	21.4	85.8	24.4	5.02	20.049	179.509	191.8	185.9
7.	7.38	50	15.8	83.3	20.5	5.36	25.046	185.506	191.1	189.3

Figure 5 Random Forest Machine Learning Model Predicted Results

3. DISCUSSIONS

The results show that RF regression is a highly effective algorithm for sugarcane crop yield prediction. The RF model outperformed the ML Regression model and Decision Tree Regression with a higher prediction accuracy. Random forest has a value of 94.6% which indicates that the predicted and the actual values of the yields are very close to each other, Multiple Linear Regression has a value of 76.5% which indicates that the predicted and actual values are fairly close to each other and Decision Tree Regression model has a value of 89.2% which indicates that the actual and the predicted values are close to each other. The models perform better compared to the observation

method which gives an accuracy of 28% and the Statistical method which gives 59% accuracy.

Even though RF has widely been used for classification in recent studies, to date few studies have used its regression capability for yield and productivity studies in ecology and agriculture. The results in this research indicate that Random Forest Regression is desirable for prediction in the agricultural field.

RF algorithm intrinsically separates random subsets of the dataset to perform testing using the remaining dataset for training the model. It also provides important information about the variables and the dependence.

The results showcase the utilities of Random Forest Regression for sugarcane crop yield prediction. There are many advantages over traditional ML Regression and Decision Tree Regression in prediction. Random Forest regression models have been shown to perform better than MLR and DTR in the explanation of variability in the dataset and the results strongly indicate the case in sugarcane crop yield prediction. It also has an upper hand when the predictors of sugarcane yield highly correlate with each other. RF regression chooses the best variable after splitting the responses at each node of the trees and averages the trees of the forest. RF can also use many types of predictor variables more easily than traditional linear and non-linear regressions. For example: Area (Ha), Age, Rainfall, Temperature, Soil PH, Soil Silt, soil k, and Yield are continuous while Crop Cycle is categorical.

The dataset was not easy to prepare because of many factors which include, mistrust from the sugarcane company, improper storage of data, noise in the data, many unnecessary features of the data, and many others. This made the research so hard.

The percentage improvement of the new model is;

Old model percentage = 59%

New model accuracy = 94.6%

= (Difference of old and new model accuracy / old model accuracy) * 100%

= (35.6/59) *100%

= 60.3%

4. CONCLUSION

The project presents a machine learning prediction model for sugarcane crop yield prediction. It allows real-time prediction throughout the year, also currently applicable in Uganda at Kakira Sugar Limited in Jinja.

The Machine Learning model was developed to predict sugarcane crop yields prediction and examined several factors that affect model performance. It provides us with state-of-the-art accuracy in prediction and shall have a great role in sugarcane production. Data science methodology

https://doi.org/10.59568/JASIC-2023-4-1-03

was used and the design of the models was done. Three Machine learning algorithms were used to implement the model Design with the sugarcane dataset. The three regression models have a good performance on the dataset provided to them. Out of them, Random Forest has the best performance with an accuracy of 94.6% which can make an accurate prediction of the sugarcane yield. The prediction model can be used in sugarcane industries, organizations, and the farmers of sugarcane. The percentage improvement of the new model is 60.3% compared to the statistical method.

5. REFERENCES

[1] Dimo Dimova, J. H. (2022). Sugarcane yield estimation through remote sensing time series and. ELSEVIER B. V, 13.

[2] Herbet, I. (2020). Prediction of sugarcane at Kakira Sugar Limited. Logallo, N. (2019). Data Science Methodology 101. Towards Data Science.

[3] Marsland, S. (2009). machine learning An Algorithm Perspective. Massey University.

[4] S R Krishan Priya a, R. K. (2022). Sugarcane yield forecast using weather based discriminant analysis. ELSEVIER B. V, 4.

[5] Spencer, R. (February 21, 2020). The Sugar Series : Sugar Cane Production – Growing, Harvesting, Processing and Refinement. CZARNIKOW.

[6] Terence Epule Epule1, J. D. (2018). The determinants of crop yields in Uganda: what is the role of climatic and non-climatic factors? Agriculture & Food Security.

https://doi.org/10.59568/JASIC-2023-4-1-03