MACHINE LEARNING MODEL FOR RICE YIELD PREDICTION USING KNN REGRESSION

Akhil Wilson, Raji Sukumar and *Hemalatha N

akhilwilson999@gmail.com, , rajivinod.a@gmail.com, hemalatha@staloysius.ac.in*

Abstract— The prediction of agriculture yield is the one of the challenging problem in smart farming,we have predicted the yield of rice in the state of Kerala, India with the help of Machine Learning by considering the soil properties, micro climatic condition and area of the rice. Here we have used Decision Tree Regression, Random Forest Regression, Linear Regression,K Nearest Neighbour Regression,Xgboost Regression and Support Vector Regression algorithms in order to predict the rice yield. From the experiments we got KNN regression to be the best with 98.77% accuracy.

Keywords:Agriculture,Preprocessing,Analysis,Regression,prediction

I. INTRODUCTION

The data science helps to predict the crop yield before harvest. In Machine Learning the models are need to train with sufficient observation. The output is predicted from the historical data. The Aim of implementing a Machine Learning model in yield prediction is nothing but it helps to reduce the difference between the actual and expected yield. If we could reduce this difference, we can reduce the farmer suicides to a certain extent. The yield prediction, just by using the past year yield may not be a good practice. We need to consider the present climatic conditions and soil properties too. Choosing a right algorithm and handling the volume of the data is challenging in ML. Here we are analyzing the past yield and predicting the future yield of the rice in the state of Kerala with the help of ML by considering certain agriculture parameters that affects the crop yield.

In this paper we are analyzing the past yield and predicting the future yield of the rice in the state of Kerala with the help of ML by considering certain agriculture parameters that affects the crop yield. Paper has been divided into three sections. Section 2 describes the methodology followed in the work. Results and discussions are described in section 3, followed by conclusion in Section 4.

II. METHODOLOGY

A. Data Source

The data set is collected from the Department of Soil Survey and Soil Conservation Government of Kerala, Planning and Economic Affairs Department Government of Kerala, Kerala agriculture university, Meteorological department Government of India. It is collected by Focus Group Discussion, Documents, Records, Oral histories and surveys. The data studies the Area and the Yield in all the block panchayats in the state of Kerala, India..The whole research is about the crop rice.The data contains the observations from all 152 block panchayats in Kerala.

B. Features Used

The data has features like District, Bloacks, Soil Types, Organic Carbon,Phosphorous, Pottassium, Manganese,Boron, Copper, Iron, Sulphur, Zinc, Soil PH, Temperature, Humidity, Precipitation, Crop, Area, Past Yield. The Data in xlsx format. The whole data is collected from all the block panchayats in Kerala. The structure of the data is shown below in Table1.

Variable Name	Variable Type	Description
State	Categorical	Kerala state only
District	Categorical	All the districts in Kerala
Blocks	Categorical	All the blocks panchayat in Kerala state
Soil Types	Categorical	Types of soil in Kerala
Organic Carbon(%)	Numeric	A soil component
Phosphorous(Kg/Ha)	Numeric	A soil component
Potassium(Kg/Ha	Numeric	A soil component
Manganese(ppm)	Numeric	A soil component
Boron(ppm)	Numeric	A soil component
Copper(ppm)	Numeric	A soil component
Iron(ppm)	Numeric	A soil component
Sulphur(ppm)	Numeric	A soil component
Zinc(ppm)	Numeric	A soil component
Soil PH	Numeric	A soil component
Temperature(deg.celsius)	Numeric	Micro climate
Humidity	Numeric	Micro climate
Precipitation(in)	Numeric	Micro climate
Crop	Numeric	Types of the crop
Area(Ha)	Numeric	Area of farming
Yield(Tones)	Numeric	Quantity obtained from give area

Table1:Structure of data

C. Work Flow

The different process in the ML model involves the study of the current farm management systems. It involve Granular,Agrivi,Trimble,FarmERP are most commonly used farm management applications. RML Farmer, pusa Krishi, AgriApp, Khethi-badi, Krishi Gyan, AgriMarket these are some of the other Indian Agricultural applications. For this research work the data is collected from the various government sources of Kerala. The feature involves micro climatic conditions, Soil properties and area of the paddy. Then the data cleaning process which involve missing value detection, Skewness, outliers are treated. Using The standardisation and normalization techniques the data transformations applied to the dataset. By using the principal component analysis, the data reduction performed on the dataset. The data is splits into training and testing then after the successive splitting of the data the different regression algorithms such as Linear Regression,K Nearest Neighbour Regression,Xgboost Regression and Support Vector Regression algorithms applied in order to predict the rice yield. Based on accuracies obtained from different models, the best algorithm is selected for the model deployment.

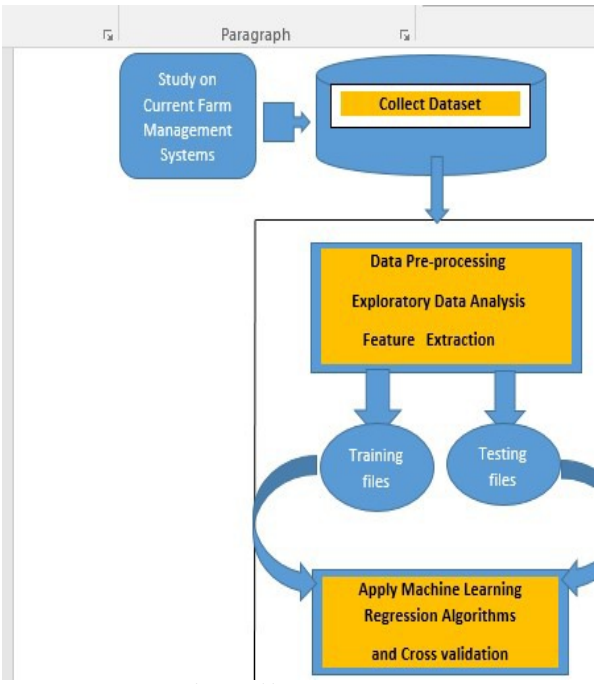

Fig 1:Architecture

D. Data Preprocessing

a) Checking Missing Values

One of the major problem that ML is facing that the presence of the missing values in the data. We have to detect the missing values in the dataset and we cannot simply ignore the missing data's in the dataset. We need to treat them. There are different deletion and imputation techniques available in order to treat the missing values in the dataset. Here in the rice dataset we found that there is no missing values associated with this particular dataset. So we need no treatment on this.

b) Checking Skewness

As simple as that skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. The skewness does not detect the outliers but it can give the direction to the outlier in the dataset. The data can be positively skewed, negatively skewed or normally distributed. We can use boxplots and distplots to find the skewness in the data. If Q1 is lower quartile, Q2 is the median and Q3 is upper quartile then the positive skewness can be shown by measuring the distance between the quartiles that is

Q3-Q2>Q2-Q1 eq(1)

By using the quartile distance we can also determine the negative skewness. The negative skewness exists when

Q3-Q2<Q2-Q1 eq(2)

Similarly, the data data is symmetrically distributed when

$$Q3 - Q2 = Q2 - Q1 \quad \text{eq(3)}$$

In this rice dataset there exist skewness, so we need to treat the skewness. There is different treatment available for skewness which involve log transformation, Box-cox transformation etc. Here in this particular project we have used the box-cox transformation in order to treat the skewness. In box-cox transformation λ which is varying from -5 to 5. Here all λ values are considered and the optimal value for the λ is selected. The transformation of the y can be represented as

$$y(\lambda) = \begin{cases} \frac{y^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log y, & if\ \lambda = 0 \end{cases} \quad \text{eq(4)}$$

This we can use for the positive data only. similarly, we can apply the box-cox transformation to the negative values by applying the following formula.

$$y(\lambda) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, & if\ \lambda \neq 0 \\ \log(y + \lambda_2), & if\ \lambda = 0 \end{cases} \quad \text{eq(5)}$$

c)Checking Outliers

Outliers are nothing but the observation which is different from the other observations in the data. With the help of the boxplots we can easily detect the outliers exists in the dataset. Usually we can say that the outliers are associated with the Inter Quartile Range(IQR).This shows that how the data is spread about the median.

$$IQR = Q3 \quad Q1 \quad \text{eq(6)}$$

where Q1-First Quartile and Q3-Third Quartile.

The inter quartile rule can be used to detect the outliers. The rule involves the detection of upper bound and the lower bound from the data. The observations which is not between the lower and upper bound is considered as the outliers.

$$Lower\ Bound = Q1 - 1.5 \times (IQR) \quad \text{eq(7)}$$

$$Upper\ Bound = Q3 + 1.5 \times (IQR) \quad \text{eq(8)}$$

Here in the rice dataset which is used here for the ML studies consists of outliers which is detected with the help of boxplots. So we need to treat the outliers here. There are different techniques available for outlier treatment such as outlier dropping method and Winsorizing methods. Here we have used the winzorising method to treat the existing outliers in the dataset. We have created a python function in order to apply Winzorisation to the outliers. This technique replaces the outlier values between the upper bound and lower bound.

d) Data Transformation

For data transformation here we have used min max normalisation.Since the measurement of the features are different we need to apply the normalization techniques in this particular research work.In this technique the minimum value is converted into zero and the maximum value is converted into one.All other values in the dataset is converted into a decimal between zero and one.we can say that

x_(scaled=(x-x_minimum)/(x_maximum) eq(9)

x -the observations in the data

x_minimum - the minimum value in the observation

$x_{maximum}$ - The maximum value in the observation

e) Data Reduction

The dataset has 20 features so we can apply the data reduction techniques here. Here we applied the principal component analysis as a dimensionality reduction technique. The principal component analysis basically converts large set of features to smaller sets that still consists of most of the information in the data.

f) Label Encoding

The dataset contains four non numerical variables such as State,Block,District,Soil Type.In order to apply the machine learning algorithm we need to convert these non-numerical feature into numerical form. we replace the categorical value with a numeric value between 0 and the number of classes minus 1. In python the LabelEncoder encodes the labels with a value between 0 and n_classes-1 where n is the number of distinct labels. When the labels getting repeated it will assign the same value to the labels as assigned earlier. Here the categorical data is converted into numerical form. Here LabeEncoder() package to encode the labels which have text data.

E . Data Splitting

In order to apply the machine learning algorithms to predict the rice yield we need to split the data into train and test sets. Here for this particular research work the data splits in the ration of 80:20. After the successful splitting we can apply the different supervised algorithms to predict the rice yield.

E. Algorithms

a) Linear Regression

Linear regression is used to represent the relationship between two variables by applying a linear equation to the observed data. Here one variable is supposed to be an independent variable, and the other is to be a dependent variable.We can say that the measure of the extent of the relationship between two variables is shown by the correlation coefficient. The range of this coefficient lies between -1 to +1. The coefficient shows strength of the association of observed data for two variables. A linear regression line equation is written in the form of:

$$y = ax + b \qquad \text{eq(10)}$$

b) Decision tree regression

Here in the decision tree regression multiple regression trees are created from the dataset and then from these trees it will predict the future yield. With the help of a set of questions the decision tree is created. Based on the questions and the answers the decision nodes are created. By using MSE (mean squared error) value the nodes splits into sub nodes. Here we used decision regressor to predict the rice and pepper yield

c) *Random Forest Reression*

Random Forest is one of the best algorithm that are used for Machine Learning studies .Here we used random forest algorithm for regression studies.One of the advantage of this algorithm is it predicts outcomes with higher accuracy most of the times even when data sets that does not have proper parameter tuning. Therefore,we can say that it has a simplicity compared to the other algorithms and it is very much popular. In the Random Forest it creates a forest in random manner. Here Multiple decision trees area created with this algorithm and are merged together in order to produce even more accurate predictions. From this algorithm clearly we can say that, as the number of decision tree in this algorithm increases the stability of the predictions also increasing.Random forest itself is working as an ensemble algorithm hence it will give good accuracy in most of problems.

d)KNN Regession

K nearest neighbors is a simple algorithm which stores all available cases and predict the numerical target based on the similarity measure. A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. We can say that the KNN regression uses the same distance functions as KNN classification. Here this algorithms is basically works on the basis of distance fuctions.We can set the number of neighbors in this algorithm. Here in this particular project we have used python environment to implement this algorithm

e)XGBoost Regression

In supervised regression algorithms XGBoost algorithm has an importanat role.Ensemble learning process involves the training and the combining of individual models to get a single prediction out of it, and we can say that this XGBoost algorithm is one of the ensemble algorithm. Here multiple trees are created and each tree learns from the previous tree. Therefore, the xgboost algorithms has an advantage such that each tree learns the updates residuals.

f) Support Vector Regression(SVR)

In the regression analysis we cannot ignore the support vector regression algorithm because for this type of regression problems this algorithm fits well. SVR is a supervised machine learning technique.SVR The support vector machine(SVM) can be used for both classification and regression but here we are used support vector regression algorithm.There are some small differences betweenthe SVM and SVR algorithms. IN SVR a hyper plane are the marginal planes are created in order to minimize the error. In support vector algorithm we increase the marginal distance to get accurate predictions. When it is the case of regression, there is a margin of tolerance which is set for the approximation. However, the main idea is always the same, to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

III. RESULTS AND DISCUSSIONS

A. Pre-processing results

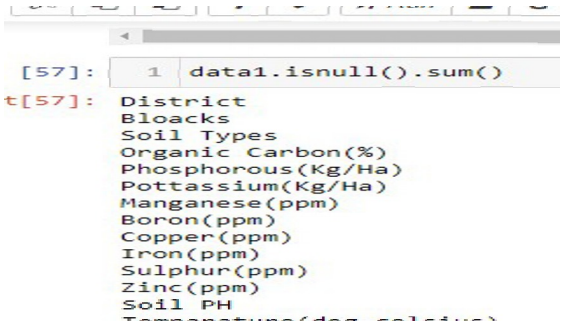

Fig2: Missing Values

Here there is no missing values associated with the datasetskewness for rice dataset can be visualize graphically as shown below.

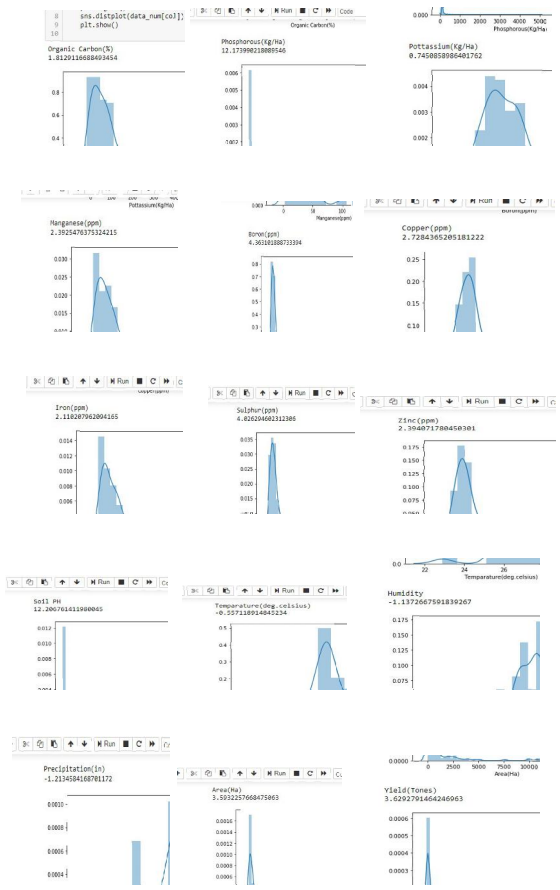

Fig 3 :distplot

Using the distlplot the skewness represented.It is found that there is skewness exist in the dataset.so we treated the skewness using the box cox transformation.the skewness after the box-cox transformation is given below

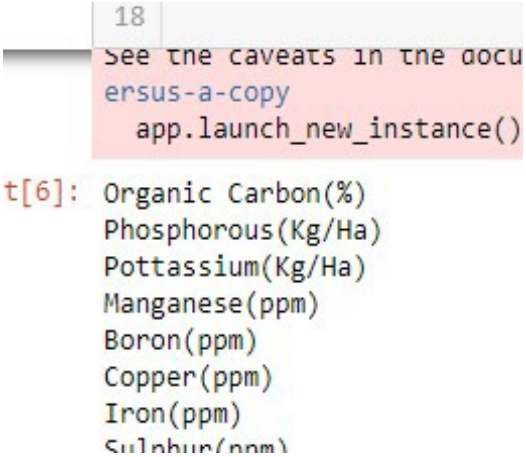

Fig 4:Skewness after the box-cox transformation

The outliers are checked with help of boxplots and here in this dataset there exist outliers and itshown below

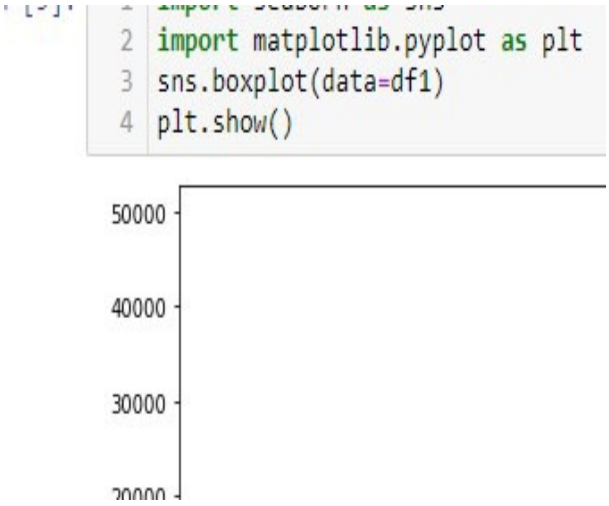

Fig 5: boxplot to detect outliers

The outliers are treated using the winsorizing method and the result is shown below.The outliers was present in the four variables and it is suceefully treated.

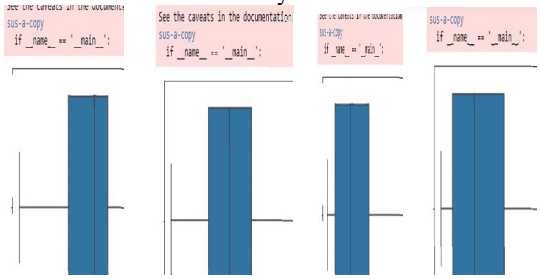

Fig 6: boxplot after winsorizing

In order to normalize the data we have used min max normalization and the result is given below

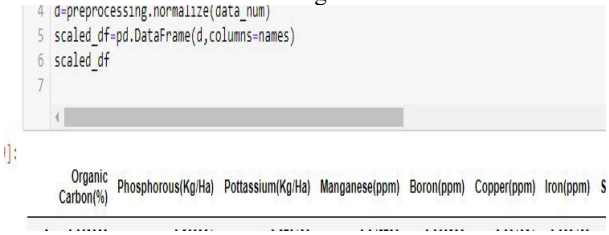

Table 2:Normalized

The data reduction is performed by using the principal componenet analysis and the result is given below

Explained variation per principal component: [0.23492603 , 0.16536821 ,0.11506274 ,0.10910176]

Here the principal componenet1 has 23.49% of information, the principal component 2 holds 16.53% of information, the principal component 3 holds 11.50% of information, the principal component 4 holds 10.91% of information.

3.2Visualization results

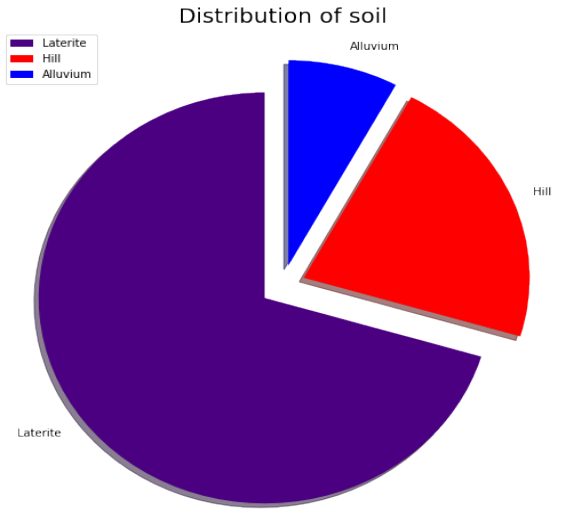

Fig7: Pie chart-soil types

Here we can see that Laterite soil is present most of the places in Kerala and all most of the crop can be farm in this type of soil. The contents in this soil support all the crops not only rice but also other crops in Kerala .

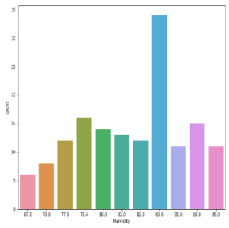

Fig8 Humidity

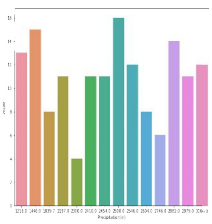

Fig9: precipitation

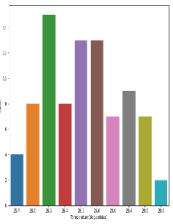

Fig10: Temperature

83.0 is the average Humidity occurs in Kerala when comparing the Humidity's in all the districts in Kerala. 2500(mm) is the most occurring precipitation in entire Kerala.It is found that 26.2 degere Celsius temperature is the most occurring temperature in kerala which is suitable for rice.

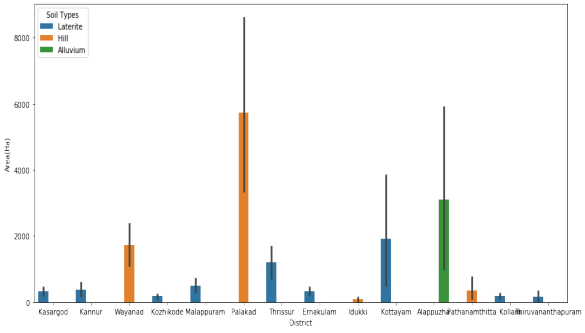

Fig11: Bar chart of yield vs Area(Ha)

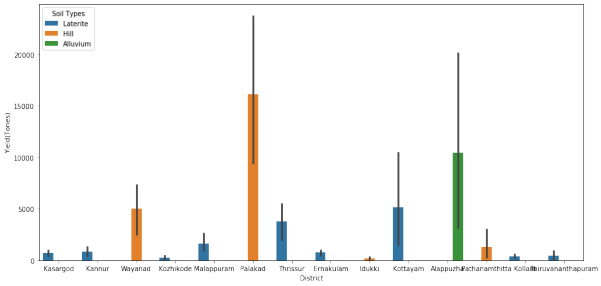

Fig12: bar chart district yield for rice

It is found that the high yield of rice obtained in Hill soil and alluvium soil. Here from the plot it is clear that the yield is high in Palakkad and Alappuzha. It is found that in Hill soil and Alluvium soil the crop yield of rice is high. The area that rice is cultivated with laterite soil is very less. The Alluvium soil is mostly present in the Alappuzha district and the 9 districts have cultivated the rice in laterite soil but the yield is less for this type of soil. From the graph it is found that the presence of the hill soil makes an impact on rice production. In Kozhikode and Idukki, the yield of the rice is very less when comparing with the other 12 districts.

District
Kasargod
Kannur
Kozhikode
Wayanad
Malappuram
Palakad
Thrissur
Ernakulam
Idukki
Kottayam

Table3: District vs Area of rice

District	Yield
Kasargod	4401.444
Kannur	9077.418
Kozhikode	3336.563
Wayanad	20127.998
Malappuram	24776.735
Palakad	209505.165
Thrissur	59972.328
Ernakulam	10501.827
Idukki	1510.249
Kottayam	57038.677
Alappuzha	125115.538
Pathanamthitta	10478.868
Kollam	4433.042
Thiruvanathapuram	4916.22

Table4: District vs Yield of rice

The important observation that we got from this is the palakad is getting high yield for rice but when comparing with the other districts it is found that the palakad having the area for rice crop cultivation is very less.The soil and climatic conditions of palakad helps to produce high yield of rice even in the less area of farming.When comparing the alappuzha district with the palakad it is found that iven the alappuzha has more area of farm land for rice it is giving less yield than the palakad.The idukki district using very small area for rice crop and the yield obtained from this district is very less.A total of 545192.072 Tonnes of rice obtained from 152 block panchayats in Kerala in the year 2018-2019 from the 114715.4 Hector of land. So we can say that the state of Kerala has a very good role in the rice production in India. From this past data of yield, we can predict the future yield of rice crop based on the soil, climatic factors and the area of farm using the Machine learning model. The alluvium soil has a very good impact on the yield of rice in the district Alappuzha

3.3 Algorithm results

Models	Accuracy(Rice) in %
Linear Regression	95.45
Decision Tree Regression	93.94
Random Forest Regression	94.70
KNN Regression	98.77
XGBOOST Regression	94.48
Support Vector Regression	97.68

Table5: Results of algorithm

These are the different Machine laearning regression algorithms used in this particular research.There fore we can say that using these algorithm we can predict the fututre yield of a crop,Here we can predict the future yield of the riceThe different models predicted the future yield of rice with a good accuracy. The rice dataset giving the highest accuracy for KNN regression and support vector regression algorithms. Out of thses 6 models we can fix a best model for the deployment of the model.

IV. CONCLUSION

The rice yield is high in the district of palakad.A total of 545192.072 Tonnes of rice obtained from 152 block panchayats in Kerala in the year 2018-2019 from the 114715.4 Hector of land. By using the machine learning techniques we can analyses the past data and we are draw a trend out of them. Based on the past data obtained from the crops, present climatic conditions and soil properties we can predict the future yield of the crop with the help of machine learning. This will help the farmers to take appropriate decisions and precautions in the farm land. Here in this research work all the yield prediction is based on the mathematical and statistical concepts.

V. ACKNOWLEDGEMENT

The authors extend their appreciation to the Deputyship of RESEARCH AND INNOVATION wing of TECHTERN Pvt. Ltd. through SMART-AGRO research and for providing all support for this research work with the project number TTRD-DS-05-2021. This is also an extension of a post-doctoral research program under Kannur University.

REFERENCES

[1] Hegde, Niranjan G., et al. "Survey paper on agriculture yield prediction tool using machine learning." *Int. J* 5 (2017): 36-39.

[2] Priya, P., U. Muthaiah, and M. Balamurugan. "Predicting yield of the crop using machine learning algorithm." *International Journal of Engineering Sciences & Research Technology* 7.1 (2018): 1-7.

[3] Hassan, Fazal Mahmud, et al. *Agricultural yield and profit prediction using data analysis techniques*. Diss. 2018.

[4] Sivanand, K., M. Sai Amrutha, and J. Chaitra Nayak. "Web Application Development for Site Specific Crop Prediction using Machine Learning." *International Journal of Modern Agriculture* 10.2 (2021): 3538-3549.

[5] Kumar, Arun, Naveen Kumar, and Vishal Vats. "Efficient crop yield prediction using machine learning algorithms." *International Research Journal of Engineering and Technology* 5.06 (2018): 3151-3159.

[6] Bondre, Devdatta A., and Santosh Mahagaonkar. "Prediction of crop yield and fertilizer recommendation using machine learning algorithms." *International Journal of Engineering Applied Sciences and Technology* 4.5 (2019): 371-376.