PREDICTIVE ANALYSIS OF SOYBEAN CROP YIELD

Ms. Prerana R[1], Ms. Shreya S Pai[2], Ms. Yashaswini M[3], Mr. Pradeep K R[4]
Students of K.S. Institute of Technology B.E,
[4]Asst. Prof Department of C.S.E
K. S. Institute of Technology, Bengaluru-62, India

Abstract: Outlier detection in machine learning identifies events or actions that do not meet expectations of existing events in a dataset. These events can lead to various defects, frauds or errors. Outlier detection is an important issue that is researched in several research areas and also in various application domains [1]. Such values in an agriculture dataset may have effects on plant growth. So, outlier removal is an important process. Soybean being one of the most grown crops and even most asked crops in the world is contributing to 25% of world's edible oil. Outliers present in the factors such as seed, leaves, stem, seed size, roots etc. may have hostile effects on plant growth of soybean crop. Hence detection of outliers in soybean crop growth becomes an important to achieve high yields. This allows the farmers to minimize yield losses taking correct actions when necessary. The proposed system uses machine learning techniques such as linear regression and Random Forest to find anomalies and remove them, thus increasing prediction accuracy of plant growth.

Keywords- **Outliers, Soybean, regression, classification**

I. INTRODUCTION

Soybean crop is considered to be one of the most important crops in the world. The importance is not only as oil seed crop and feed for livestock but also is a good source of protein for human diet. The demand for soybean has increased since a decade which has challenged its supply. In order to meet the demand, it is important to increase the crop yield [2]. Exploitation of Soybean crop in India started four decades ago and since then the production and demand for the crop has increased unparallelly. Soybean and their derivatives are most traded commodity and accounts for about 10% of global agriculture trade. The demand for Soybean and its products has rapidly increased since 1990s and has crossed the trade for wheat and other coarse grains [3]. However, various factors may have direct or indirect effect on soybean crop growth rate. These factors include month, precipitation, temperature, hail, germination, seed, seed-size, leaves etc. Any anomalies detected in any of these attributes may delay plant growth. Thus, removal of such anomalies becomes important.

Outliers are extreme values that fall beyond other observations. Outlier detection or anomaly detection is the process of identifying those values that fall beyond the normal distribution. In agriculture, a lot of research is carried out on yield data due to its importance in crop management. Most of the datasets available in the repository may have a minimum of 10% erroneous, missing or not available values [4]. Agriculture dataset like soybean have several errors like outliers and missing values due to some unknown sources which lead to wrong prediction. For example, poor weather condition during a particular season may affect crop growth during that season. A vast number of unsupervised, semi supervised and supervised algorithms are found in the literature for outlier detection. These algorithms further can be classified to classification-based, clustering-based, nearest neighbour based, density based, information theory based, spectral decomposition based, visualization based, depth based and signal processing-based techniques [4]. The proposed model uses data pre-processing technique such as Z-score for outlier detection.

II. RELATED WORK

Varun Chandola [1] provided a structured and comprehensive outline of research on outlier detection. Here existing system is divided into different categories, and to each category key assumptions were applied to distinguish normal and abnormal behaviour. Acu˜na [5] proposed various techniques to detect outliers and also provided experimental results that show improved effectiveness of performance of classifiers on outlier removal. Nedunchezhian [7] presented missing value problems in data mining and evaluated few methods used for missing value imputation. Sánchez [11] compared predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Various algorithms were used for massive crop yield prediction in agricultural planning.

III. BACKGROUND PROBLEM

Though, previous researches have proved that outliers decrease the classifier accuracy, there are no paper suggesting the best algorithm suitable for the soybean dataset to increase the production growth. This research gap has motivated to propose this model.

IV. RESEARCH METHODOLOGY

The proposed model is evaluated using Python on soybean dataset collected from UCI Machine Learning Repository [12] to test the accuracy of the data with outliers and without it. Below subsections deal with the dataset used, experimental setup of the proposed model and results. The impacts of how classifier accuracy is increased after the removal of outliers.

PROPOSED METHOD

The proposed model is developed in 2 stages. In the first stage, the dataset is cleaned as they tend to have incomplete data and outliers which can lead to wrong prediction. Instances having missing values in more than 5 attributes are eliminated then outliers are identified using Z-score[5]. Attribute having Z-score values with a threshold greater than 3 are replaced with the mean value of that attribute. In the second stage the classification technique is applied to the pre-processed data to predict if the plant growth of soybean plant is normal or abnormal.

Data Pre-processing

Usually, data contain missing values, noisy instances and outliers as they are gathered from many sources and are integrated. Quality of the mining process depends on the quality of the data. The mining results are misled due to presence of missing values, redundant attributes and noisy instances. The proven method to overcome is data pre-processing [7].

Treating missing Values

In an instance, if no data is present for an attribute then we treat it as a missing value. It occurs due to various reasons like manual data entry procedures, unable to collect an observation, equipment errors, incorrect measurements etc. If the missing values are not handled properly then algorithm performance reduces. To overcome this problem there are 4 popular ways to handle [6][7]. They are:

1. When compared to the data if the missing value percentage is small then we eliminate it.
2. In a dataset a variable is deleted if missing values compared to other variables are more.
3. Imputation for numeric value by mean and median, and mode for categorical value.
4. Predicting the missing values by decision tree, regression, Knn etc.

The author [7] has discussed various methods to handle missing values, hence in this proposed model the first step is to eliminate the missing values present in the datasets. Among various methods to handle missing values we choose to eliminate it as it constitutes less percentage of data which is negligible.

Outlier elimination

An outlier is a specific data point that falls outside the range of probability for a dataset. In other words, the outlier is distinct from other surrounding data points in a particular way. The main aim is to find data patterns that do not fit into expected behaviour. In many database applications, the amount of fault data reaches more than ten percentage [9]. Therefore, in order to improve the quality of stored data, removing or replacing the outliers enhances.

A Z-score is a numerical measurement used for statistical estimates of a value relationship to the mean of a group of values, measured in terms of deviations from the mean [8]. After calculating the threshold of the dataset using Z-score, instances which have threshold greater than 3 are replaced by the mean value of the particular attributes.

Classification

After the pre-processing the dataset is now trained with 80-20 split with 80% training data and 20% testing data using logistic regression and random forest. The base classifier for the Random Forest is Decision Tree. It's an ensemble learning method and consists of number of trained classifiers. When a new instance is to be classified it uses these trained predictors to classify [10]. Logistic regression is a classic predictive modelling algorithm and implemented when the class variable is binary categorical [11].

FLOW OF THE PROPOSED SYSTEM

The flow of the proposed system is as shown below. The Soybean dataset is pre-processed to eliminate missing values then the irrelevant data that deviate from the normal values are identified by Z-score and are replaced by the mean value of the attribute, as they hinder analysis process. To this cleaned data Logistic Regression and Random Forest classifiers are applied to predict if the plant growth normal or abnormal.

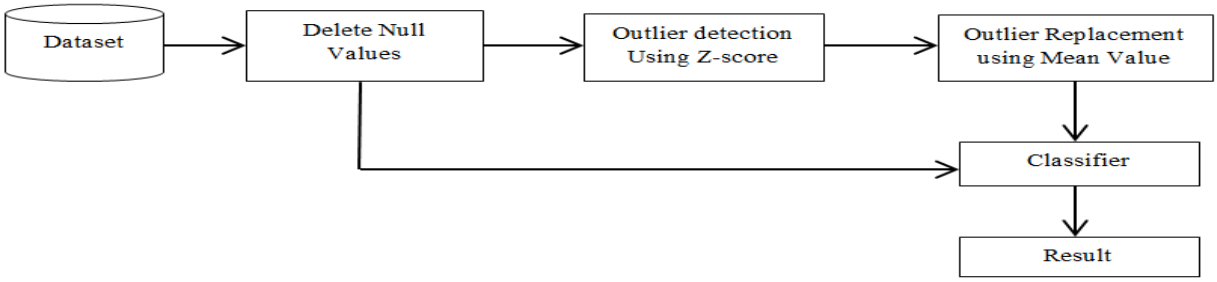


Figure 1: Flow of the proposed system

Dataset Description

TABLE 1: DATASET DESCRIPTION

Data Set	No. of Attributes Including class	No. of Classes	No. of Instances	Missing Values
Soybean	35	2	667	Yes

ESTIMATIONS FOR MODEL PERFORMANCE

Performance measures

Supervised Machine Learning (ML) has several ways of evaluating the performance of the classifiers. The quality of classification algorithms is measured based on the confusion matrix which records correctly and incorrectly recognized examples for each class. Table 4 presents a confusion matrix for binary classification, where TP are true positive TN are true negative, FP false positive, FN false Negative. The different measures used with the confusion matrix are

TABLE 2: CONFUSION MATRIX

	Predicted class		
		Test Negative (T-)	Test Positive (T+)
Actual class	Disease Absent (D-)	True Negative (TN)	False Positive (FP)
	Disease Present (D+)	False Negative (FN))	True Positive (TP)

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.
Accuracy = (TP + TN) / (TP + TN + FP + FN)

Sensitivity: Sensitivity is also referred as True positive rate i.e., the proportion of positive tuples that are correctly identified.
Sensitivity = TP/ (TP+FN)

Specificity: Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified.
Specificity= TN/ (TN + FP)

Positive Predictive Value: Precision also known as positive predictive values PPV is the proportion of the predicted positive cases that were correct.
PPV= TP/ (TP + FP)

Negative Predictive Value: Negative predictive values NPV are the proportion of the predicted negative cases that were correct.
NPV= TN/ (TN + FN)

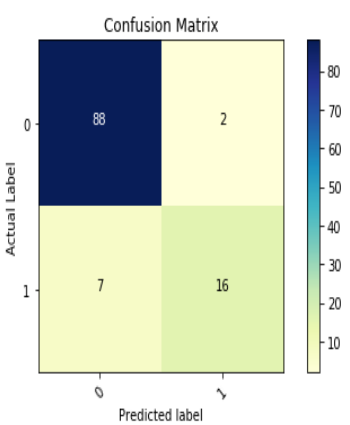
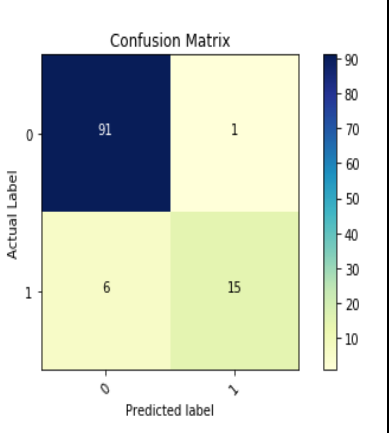
False positive rate: The false positive (FP) rate also known as Type I error is the proportion of negative cases that were incorrectly classified as positive.
FPR= FP/ (FP+TN)

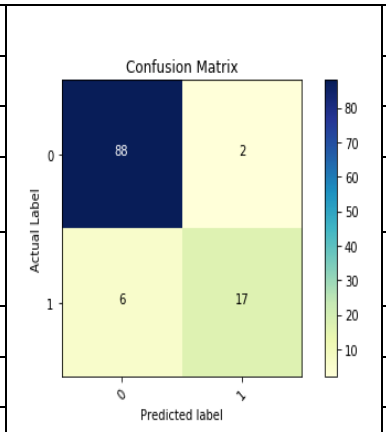
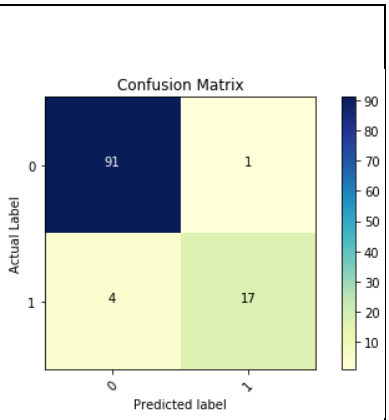
False negative rate: The false negative (FN) rate also known as Type II error is the proportion of positive cases that were incorrectly classified as negative.
FNR= FN/ (FN+TP)

False Discovery rate: FDR measures the proportion of discoveries that are false among all discoveries, i.e., the chance of not having the condition among those that test positive.
FDR=1-PPV

V. EXPERIMENTAL RESULTS

Table 3: Experimental result to predict soybean plant growth

Result table					
Classifier	**Performance Estimators**	**With outliers**	**Confusion Matrix**	**Outlier replacement**	**Confusion Matrix**
Logistic Regression	**Accuracy**	92.0354		93.8053	
	Sensitivity	0.69565		0.71429	
	Specificity	0.97778		0.98913	
	Positive Predictive Value	0.88889		0.93750	
	Negative Predictive Value	0.92632		0.93814	
	False Positive Rate	0.02222		0.01087	
	False Negative Rate	0.30435		0.28571	
	False Discovery Rate	0.11111		0.06250	

Random Forest	**Accuracy**	92.9204		95.5752	
	Sensitivity	0.73913		0.80952	
	Specificity	0.97778		0.98913	
	Positive Predictive Value	0.89474		0.94444	
	Negative Predictive Value	0.93617		0.95789	
	False Positive Rate	0.02222		0.01087	
	False Negative Rate	0.26087		0.19048	
	False Discovery Rate	0.10526		0.05556	

VI. CONCLUSION

Using Soybean dataset, plant growth is predicted. Soybean data has lot of outliers, presence of which will lead to wrong prediction whereas removal tends in loss of information therefore an attempt is made in this proposed work to identify outliers and replace it with mean value. Experimental results proved that the proposed model is successful and showed increased performance after the outlier elimination.

VII. FUTURE SCOPE

Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. These techniques are used while solving machine learning problems to obtain enhanced performance for a classification or regression task. Hence in future work it will be applied to soybean dataset to further enhance the performance of the classifier.

VIII. ACKNOWLEDGEMENT

The authors thank their management and institution for providing them the resources and platform for showcasing their idea. They would also like to thank their HOD and all the lecturers and professors for motivating them into framing this research.

IX. REFERENCES

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar, July 2009, Anomaly Detection: A Survey, ACM Computing Surveys, Vol. 41, No. 3, Article 15.
[2] Sharma, P., Dupare, B. U., & Pate, R. M. (2016). Soybean improvement through research in India and socio-economic changes, Indian Council of Agricultural Research, pp.1-3.
[3] Lee, T., Tran, A., Hansen, J., & Ash, M. (2016). Major factors affecting global soybean and products trade projections (No. 1490-2016-128405).
[4] Christy, A., Gandhi, G. M., & Vaithyasubramanian, S. (2015). Cluster based outlier detection algorithm for healthcare data. Procedia Computer Science, 50, 209-215.
[5] Acuña, E., & Rodriguez, C. (2004). On detection of outliers and their effect in supervised classification. University of Puerto Rico at Mayaguez.
[6] Gimpy, M. D. R. V. (2014). Missing value imputation in multi attribute data set. International Journal of Computer Science and Information Technologies, 5(4), 1-7.
[7] Somasundaram, R. S., & Nedunchezhian, R. (2011). Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, 21(10), 14-19.
[8] Christy, A., Gandhi, G. M., & Vaithyasubramanian, S. (2015). Cluster based outlier detection algorithm for healthcare data. Procedia Computer Science, 50, 209-215.
[9] Last, M., & Kandel, A. (2001, November). Automated detection of outliers in real-world data. In Proceedings of the second international conference on intelligent technologies(pp. 292-301).
[10] Eesha Goel* , Er. Abhilasha, Random Forest: A Review, January 2017 International Journal of Advanced Research in Computer Science and Software Engineering.
[11] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.
[12] UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information science and ComputerScience.{http://www.ics.uci.edu/~mlearnMLRepository.html}1998.