IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AGRICULTURAL PRODUCT PRICE AND CROP CULTIVATION PREDICTION BASED ON DATA SCIENCE TECHNIQUE

S.Jonisha[1] Dhanush B[2] Jayanth[3] Thiruvengadam.S[4]

Assistant Professor[1] ,Student[2,3,4]

Department of Computer Science and Engineering

PERI INSTITUTE OF TECHNOLOGY

ABSTRACT

Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to these problems, the crop production is not improved which affects the agriculture economy. Hence a development of agricultural productivity is enhanced based on the plant yield prediction. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity and Entropy.

INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at

LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

LITERATURE SURVEY

This work aims to show how to manage heterogeneous information and data coming from real datasets that collect physical, biological, and sensory values. As productive companies public or private, large or small need increasing profitability with costs reduction, discovering appropriate ways to exploit data that are continuously recorded and made available can be the right choice to achieve these goals. The agricultural field is only apparently refractory to the digital technology and the "smart farm" model is increasingly widespread by exploiting the Internet of Things (IoT) paradigm applied to environmental and historical information through time-series. The focus of this study is the design and deployment of practical tasks, ranging from crop harvest forecasting to missing or wrong sensors data reconstruction, exploiting and comparing various machine learning techniques to suggest toward which direction to employ

efforts and investments

In agriculture, crop yield prediction is critical. Crop yield depends on various features including geographic, climate and biological. This research article discusses five Feature Selection (FS) algorithms namely Sequential Forward FS, Sequential Backward Elimination FS, Correlation based FS, Random Forest Variable Importance and the Variance Inflation Factor algorithm for feature selection. Data used for the analysis was drawn from secondary sources of the TamilNadu state Agriculture Department for a period of 30 years. 75% of data was used fortraining and 25% data was used fortesting. The performance of the feature selection algorithms are evaluated by Multiple Linear Regression. RMSE, MAE, R and RRMSE metrics are calculated for the feature selection algorithms. The adjusted R2 was used to find the optimum feature subset, also the time complexity of the algorithms was considered for the computation. The selected features are applied to Multilinear regression, Artificial Neural Network and M5Prime. MLR gives 85% of accuracy by using the features which are selected by SFFS algorithm.

In many supervised learning problems feature selection is important for a variety of reasons: generalization performance, running timerequirements, and constraints and interpretational issues imposed by theproblem itself. In classification problems we are given f data points Xi E ~n labeled Y E ±1 drawn from a probability distribution P(x, y). We would like to select a subset of features while preserving or improving the discriminative ability of a classifier. As a brute force search of all possible features is a combinatorial problem one needs to take into account both the quality of solution and the computational expense of any given algorithm. Support vector machines(SVMs) have been extensively used as a classification tool with a great deal of success from object recognition to classification of cancer morphologies and a variety of other areas.

Rice blast disease (RBD) is one of the most damaging crop disease for the rice in Taiwan. RBD may be widespread and cause severe losses if it is not controlled in the early stage. The goal of this research is to build an early warning mechanism for the RBD using the machine learning model under current climatic condition. Five years of climatic data (ranging from2014 to 2018) are used as candidate feature sin ourmodel, which are collected by the Taiwan government. The RBD conditions are labeled via the field observation during these years. With the climate data, we conduct the recursive feature elimination algorithmto selectthe key featuresthat have impacts on the RBD. To derive the RBDpredictionmodel, we applied the Auto-Sklearn and neural network algorithms to train the classification model. The experiment results show that the proposed model can classify the RBD conditions (whether exacerbated or relieved) with an accuracy of 72% in average. In particular,our model can achieve an accuracy of 1389% in the exacerbation case,which demonstrates the effectiveness of the proposed classification model.

The use of feature selection can improve accuracy, efficiency, applicability and under standability of a learning process and the resulting learner. For this reason, many methods of automatic feature selection have been developed. By using the modularization of feature selection process, this paper evaluates a wide spectrum of these methods and some additional ones created by combination of different search and measure modules. The evaluation identifies the most interesting methods and shows some recommendations about which feature selection method should be used under different conditions.

SYSTEM ARCHITECTURE

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.The user will give the input data of their agricultural land those input data will be pre-process. Using the past dataset and then evaluating those input data using different machine learning algorithms. And then the high accuracy of the algorithm will be given as a GUI output.

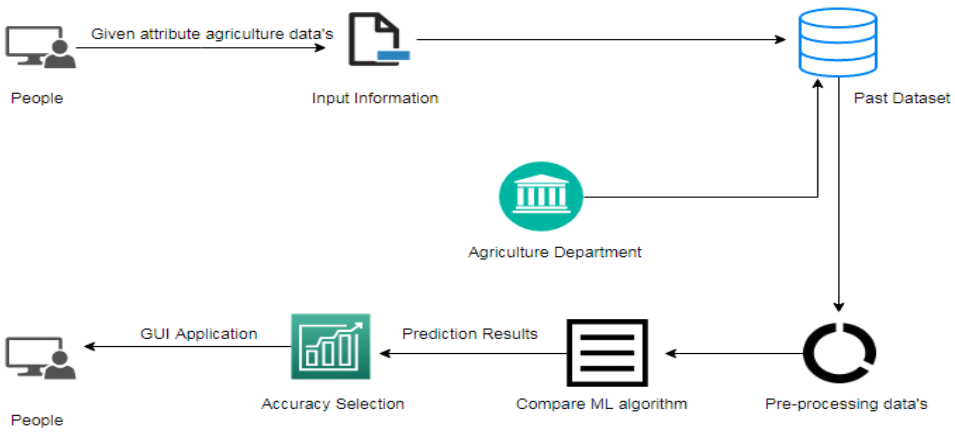

Fig 1: system architecture

PROBLEM DEFINITION

In Recent days, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to these problems, the crop production is not improved which affects the agriculture economy. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to several information's like, variable identification, uni-variate analysis, bi-variate and multi- capture variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop.

SYSTEM IMPLEMENTATION

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training datasetwhile tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop thecolumn etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Comparing Algorithm

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. A way to do this is to use different visualization methods to show the average accuracy, variance and other 28properties of the distribution of model accuracies. In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

- Random Forest
- Decision Tree Classifier
- Naive Bayes

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the crop using machine learning algorithm with different results. This brings some of the following insights about crop prediction. As maximum types of crops will be covered under this system, farmer may get to know about the crop which may never have been cultivated and lists out all possible crops, it helps the farmer in decision making of which crop to cultivate. Also, this system takes into consideration the past production of data which will help the farmer get insight into the demand and the cost of various crops in market.

Remaining SMLT algorithms will be involve to finding the best accuracy with applying to predict the crop yield and cost. Agricultural department wants to automate the detecting the yield crops from eligibility process (real time). To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment

REFERENCES

1.A.Mark Hall, "Feature selection for discrete and numeric class machine learning," Comput. S Sci., Univ. Waikato, pp. 359–366, Dec. 1999.

2. I.Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Jan. 2003.

3. R.Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection–theory and algorithms," in Proc. 21st Int. Conf. Mach. Learn. (ICML), 2004, p. 43.

4. P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 139–144, 2018.

5. S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Comput. Social Syst., vol. 8, no. 1, pp. 214–226, Feb. 2021.