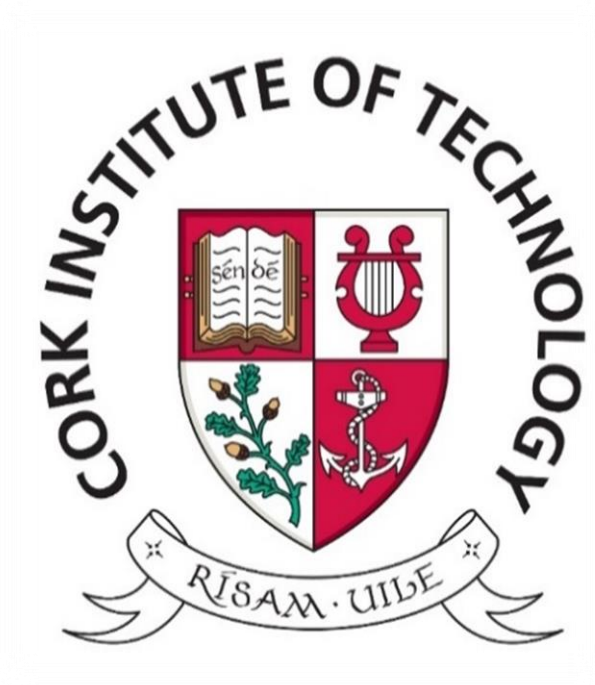
CORK INSTITUTE OF TECHNOLOGY



DEPARTMENT OF MATHEMATICS

MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS
2019-2020

CROP YIELD PREDICTION BY USING RANDOM FOREST ALGORITHM

by

SHIVAANI KATRAGADDA

(R00183214)

FOR THE MODULE MATH9001 – RESEARCH METHODS

4TH MAY 2020, MONDAY

Supervisors: ANGEUS DALY AND VINCENT CREGAN

Declaration of Authorship

I, Shivaani Katragadda, declare that this thesis titled, 'Crop Yield Prediction by using Random Forest Algorithm' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for the Masters' degree at Cork Institute of Technology.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Cork Institute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I understand that my project documentation may be stored in the library at CIT and may be referenced by others in the future.

Signed: SHIVAANI KATRAGADDA

Date: 4TH MAY 2020

TABLE OF CONTENTS

TABLES

TABLE OF FIGURES

ABSTRACT

Worldwide, agriculture is the backbone of many countries. Agriculture depends on farmer's education and knowledge to grow crops. It is also weather dependent, and it is impacted terribly by both floods and droughts. It will be beneficial for farmers to know which crop yield provides more profit. Therefore, throughout the growing and harvesting cycle of agriculture, machine learning (ML) has become popular in this multi-disciplinary agriculture domain. Machine learning is an essential approach to achieving practical and effective solutions for predicting crop yield production. Crop yield projections include estimating crop yields from available historical data such as precipitation data, soil data, and historic crop yields. Various machine learning techniques including prediction, classification, regression, and clustering can be utilized to forecast crop yield. Artificial and Deep Neural Networks (ANNs and DL), Support Vector Machines (SVMs), Ensemble learning (EL), linear and logistic regression, decision trees, Naive Bayes are some of the algorithms that can be used to implement prediction. The primary focus of this research is to use a machine learning technique to predict the yield of a crop for a particular year based on existing data by using the Random Forest algorithm. Random Forest, is one of the most powerful and popular supervised machine learning algorithms. It can be used to predict the future of crop yield accurately. The dataset entitled "Production_Crops" is sourced from the Food and Agriculture Organization of the United Nations (FAO) this data will be used for building the model. Knowledge of estimated predictions can help farmers to predict the yield of a crop before cultivating it on to the agriculture lands or fields.

Keywords: Crop Yield Prediction, Crop Analysis, Agriculture, Machine Learning, Random Forest.

WORD COUNT:278

1 INTRODUCTION

Agriculture is one of the important industrial sectors in any country because many countries economy relies mainly on agriculture for the sustainability of its rural areas. In a country, for example, India, which has ever increased demand for food due to the rising population, advances in the agriculture sector are required to meet the needs of its population. Due to poor education, many farmers in India do not have knowledge or awareness regarding the cultivation of the crops to plant and when to plant them. Due to lack of these cultivating techniques and some of which factors[1] including climate change, unpredicted rainfall patterns, falling water levels, excessive use of pesticides, and many more factors, the level of agricultural production in India is declining. Most farmers do not achieve the expected crop yield for a variety of reasons. To understand production levels, yield prediction is carried out which involves predicting the yield of the crop based on the existing data. Previously, crop yield estimates were based on farmers cultivation experience. Accurate information about the history of crop yield is an important thing for making decision-related to agricultural risk management.

1.1 PROBLEMS

Crop yield prediction is a major problem in agriculture. However, there are many ways to predict and improve crop yield and quality. Starting at each growing season, agricultural planners require estimations of the yield for all the involved crops[2]. Crop yield prediction can be difficult because it depends on many interrelated factors and moreover, the yield is also impacted by a farmer's decision (such as applied irrigations, pest and fertilizers applications, crop rotation, and land preparation) and uncontrollable factors (such as weather, subsidies, and market).

1.2 NEED FOR PREDICTION

Estimating agricultural yield is an important issue in agriculture, as the changes from year to year in crop yield influence food supply, international business, and global market prices. Also, an appropriate prediction of crop productivity is required for efficient planning of land usage and economic policy. In recent times, the forecasting of crop productivity within-field level has increased. Early prediction of crop yield provides most of the useful information to policy planners. The most influencing factor for crop productivity is weather conditions. If the weather-based prediction is made

more precise, then farmers can be alerted well in advance so the major loss can be mitigated and would be helpful for economic growth. The prediction will also aid the farmers to make decisions to choose the alternative crops at an early stage in case of critical situations. Further, predicting crop yield can facilitate the farmers to have a better vision of the cultivation of seasonal crops and their scheduling. Thus, it is necessary to simulate and predict the crop yield before cultivation for efficient crop management and the expected outcome. As there exists a non-linear relationship between crop yield and the factors influencing crop, machine learning techniques might be efficient for yield predictions.

1.3 DOMAIN

Data mining[3] is useful for predicting crop yield production. It is defined as a process of finding patterns among multiple fields in large relational databases which can be converted into historical patterns and future trends.

1.4 RESEARCH AREA

Machine learning[4] algorithms can help in improving the production of crop yield rates. Whenever there is a loss in unfavorable conditions, we can apply the crop selecting method and reduce the losses. And it can be used to gain crop yield rate in favorable conditions. This maximizing of yield rate helps in improving the country's economy.

1.5 AIM

The intention of this research is to propose an idea to predict the yield of the crop. The farmers will check the yield of the crop as per the acre, before cultivating onto the field. The main aim of this research is to implement crop yield prediction. One can predict the production of each crop by using the Random Forest classifier[5]. which will provide an estimate of the crop yielding rate. Therefore, it can be established which crops have higher demands for farming in the upcoming years.

1.6 RESEARCH QUESTION

The main objective of the study will be to answer the following questions concerning the study area:

- Does crop production correlate with the increasing population?
- Does crop yield correlate with the increasing population?

WORD COUNT: 702

2 LITERATURE REVIEW

Machine Learning deals with problems where the relationship between input and output variables is not known or hard to obtain. The "learning" term here denotes the automatic acquisition of structural descriptions from examples of what is being described. Unlike traditional statistical methods, machine learning does not make assumptions about the correct structure of the data model, which describes the data. This characteristic is very much useful to model complex non-linear behaviors, such as a function for crop yield prediction. Machine Learning techniques are successfully applied to Crop Yield Prediction. The Supervised Learning Algorithm consists of a target(dependent) variable which is to be predicted from a given set of predictors(independent). Using these sets of variables, we generate a function that map inputs to get the desired outputs. The process of training set continues until the model achieves a desired level of accuracy on the training data.

2.1 RELATED WORK

The paper, "Crop Selection Method to maximize crop yield rate using machine learning technique"[6] has concluded that this paper helps in improving the yield rate of crops can be achieved by applying classification methods and comparing the parameters. The algorithms used in this paper are Bayesian, Support Vector Machine, Clustering, K-means. However, the paper noted that the disadvantage of using these algorithms is that there is no proper accuracy and performance.

The publication, "Applications of Machine Learning Techniques in Agricultural Crop Production"[7] shows that this is an advanced researched field and is expected to grow in the future. The integration of computer science with agriculture helps in forecasting agricultural crops. This method also helps in providing information about crops and how to increase the yield rate. The algorithms used are Artificial neural networks, Decision Tree, Regression analysis, and the disadvantage is clear methodology is not specified.

The aim of the following paper "A Model for Prediction of Crop Yield" [8] is to propose and implement a rule-based system. And predict the crop yield production from the collection of previous data. The algorithms used are clustering, Kmeans and the disadvantage is Suitable only for using association rule and considered a small amount of data.

The paper named, “Machine Learning: Applications in Indian Agriculture” [9] reviews various applications of machine learning in the farming sector. And also provides an insight into the troubles faced by many Indian farmers and how these can be solved using these techniques. The algorithms used are Decision Tree, Bayesian Belief Network, Artificial neural networks, Regression analysis.

The study “Artificial neural networks for rice yield prediction in mountainous regions”[10] developed an application using the simple and accurate estimation techniques to predict rice yields in the planning process. The necessity of the study was to:

(1) Identify whether or not that an artificial neural network model could predict rice yield approximately for typical climate conditions of the mountainous region.

(2) Evaluate Artificial neural network model performance related to a variety of developmental parameters.

(3) Compare the effectiveness of more than one linear regression models with Artificial neural network models.

The paper “Crop Selection Method Based on Various Environmental Factors Using Machine Learning”[11] discusses predicting crop sequence, maximizing yield rates, and making benefits to the farmers. It used machine learning applications with agriculture in predicting crop diseases, studying crop simulations, different irrigation patterns. The algorithms used in this study were Support Vector Machine and Artificial neural networks.

The study “Analysis of crop yield prediction using Data Mining Techniques” [12] used multiple linear regression method which can be applied for existing data and hence helps in analyzing and verifying the data for the crop yield.

In “Crop yield prediction using Agro Algorithm in Hadoop”[13] the authors have concluded that this method will provide an agro algorithm which helps in predicting suitable crop for the lands. And this helps in improving the quality of the crop.

The study “Application of Data mining tool to crop management system”[14] has concluded that the regression analysis method will help in estimating rainfall and investigate the reasons for getting a lower yield.

According to the paper “Application of Machine Learning Techniques for yield prediction on delineated zones in precision agriculture” [15], three methods are used which include Clustering, Classification, and Normalization which helps in analyzing and understanding crop yield rate for zones which is based on attributes. The disadvantage is that it gives the only framework.

2.2 RANDOM FOREST CLASSIFIER

Random Forest is a most famous and powerful supervised machine learning algorithm capable of performing both classification, regression tasks, that operate by constructing multiple decision trees at training time and outputting the class that is the mode of the classification or regression of the individual trees. The more trees in a forest the more robust the prediction. Random decision forests correct the decision trees habit of overfitting to their training set.

Random Forest is a non-parametric advanced classification and regression tree (CART) analysis technique that has been adopted widely in many scientific fields, including predicting suitable habitat of various plant and animal species[16][17][18][19] and gene expression interpretation[20][21]. A majority of applications of random forest have been focused on its utility as a classification tool with only limited studies exploring its regression capabilities for predicting ecosystem or crop productivity[22][23][24]. Several studies have pointed to a number of promising advantages as well as drawbacks of the random forest as a regression tool over traditional regression models[25][26]. To date, random forest regression applications in the fields of agronomy and crop science remain scarce, with few exceptions[22].

WORD COUNT: 905

3 METHODOLOGY

The main aim of this research is to show the method to improve the crop yield prediction thus resulting in profitability to the farmers.

3.1 DESIGN FLOW FOR CROP YIELD PREDICTION

The process of developing the proposed model system involves the following process:

1. Dataset collection

2. Pre-processing

3.Feature Selection

4.Split the dataset into Train and Test set

5.Applying Machine Learning Technique

3.1.1 DATASET COLLECTION

The dataset "Production_Crops" is taken from Food and Agriculture Organization of the United Nations (FAO)[27]. The dataset consists of 51,154 entries and 123 total columns. Crop statistics are recorded for 173 products, covering the following categories: Crops Primary, Fibre Crops Primary, Cereals, Coarse Grain, Citrus Fruit, Jute Jute-like Fibres, Oil crops Primary, Treenuts, Vegetables and Melons. Data is expressed in terms of area harvested, production quantity and yield. The dataset consists of all primary crops of all countries according to the United Nations M-49 list. The attributes of the dataset are as follows:

Attribute No	Attribute Name	Attribute Description
1	Area code	Area code of all countries of the world and geographical aggregates according to the United Nations M-49 list.
2	Area	All countries of the world and geographical aggregates according to the United Nations M-49 list.
3	Item code	Item code of each crop
4	Item	Name of the Items
5	Element code	Code of each element

6	Element	Data is expressed in terms of area harvested, production quantity and yield
7	Unit	Production Quantity and Seed: tonnes; Area harvested: hectares; Yield: tonnes per hectare.
8-123	Y1961, Y1961F to Y2018, Y2018F	The year indicates that the data refer to production, yield and area harvested during the calendar year

TABLE 1: DETAIL EXPLANATION OF DATASET

3.1.2 PRE-PROCESSING

Pre-Processing consists of inserting the missing values, the appropriate data range, and extracting the functionality.

3.1.3 FEATURE EXTRACTION

It is the process of decreasing the raw data into manageable groups(features) for processing it. Beginning with a preliminary set of raw data it builds up features that result in informative and non-redundant data. As the statistical agriculture data is redundant and too large to be processed, it is first transformed into a reduced or minimal set of features.

3.1.4 SPLIT THE DATASET INTO TRAIN AND TEST SET

In this step, the loaded records are split into two sets, such as training data and test data, with a division ratio of 70% or 30%. The training set is the only one on which we train and fit our model basically to fit the parameters whereas the test data is used only to assess the overall performance of the model. Training data is available to model whereas testing data is the unseen data for which predictions have to be made.

3.1.5 APPLYING MACHINE LEARNING TECHNIQUE

In this proposed thesis the supervised machine learning algorithm Random Forest will be used for crop yield prediction which is given as follows:

3.1.5.1 RANDOM FOREST

A collection of trees is a forest, and the trees are being trained on subsets that are being selected at random, hence it is called random forest. Random Forest is an ensemble classifier. To construct random forest, a collection of decision trees are required by considering two-third of the records in the datasets. These decision trees are applied to the remaining records for the accurate

classification. The resultant training sets can be applied to the test data for the proper and correct prediction of crop yield.

This algorithm runs efficiently and has higher classification accuracy. The principal parameters in the random forest[28] algorithm are as follows:

Ntree - Number of trees to grow. Larger the tree, it will be more computationally expensive to build models.

Mtry - A variety of variables arbitrarily sampled as candidates at each split. Note that the default values are unique for classification(sqrt(p) where p is number of variables in x) and regression(p/3).

Nodesize - Minimum length of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take time). Note that the default Values are distinctive for classification and regression.

3.1.5.2 DECISION TREE

The Decision tree classifiers use a greedy method for this reason an attribute chooses at the beginning step cannot be used anymore which can give better classification if used in later steps. Also, it overfits the training data that can give poor results for unseen data. So, to overcome this problem ensemble model is used. An ensemble model results from different models are combined. The result acquired from an ensemble model is usually better than the result from any one of individual models.

3.2 TOOL USED

R is the most used tool for statistics, data analysis, and machine learning. It is a programing language; therefore, we can produce our own functions, objects, and packages. It is an open-source platform and might be used on any operating system.

3.3 RISKS

- Once conducting any variety of analyses, one should be ready for the restrictions that one may encounter and there could also be no access to data and guaranteeing that the information meets the data protection criteria below the final information Protection Regulation (GDPR).
- Another risk involves not understanding the way to interpret information so as to clarify it.
- Last however not least time constraints to complete the research.

3.4 FLOW CHART FOR CROP YIELD PREDICTION

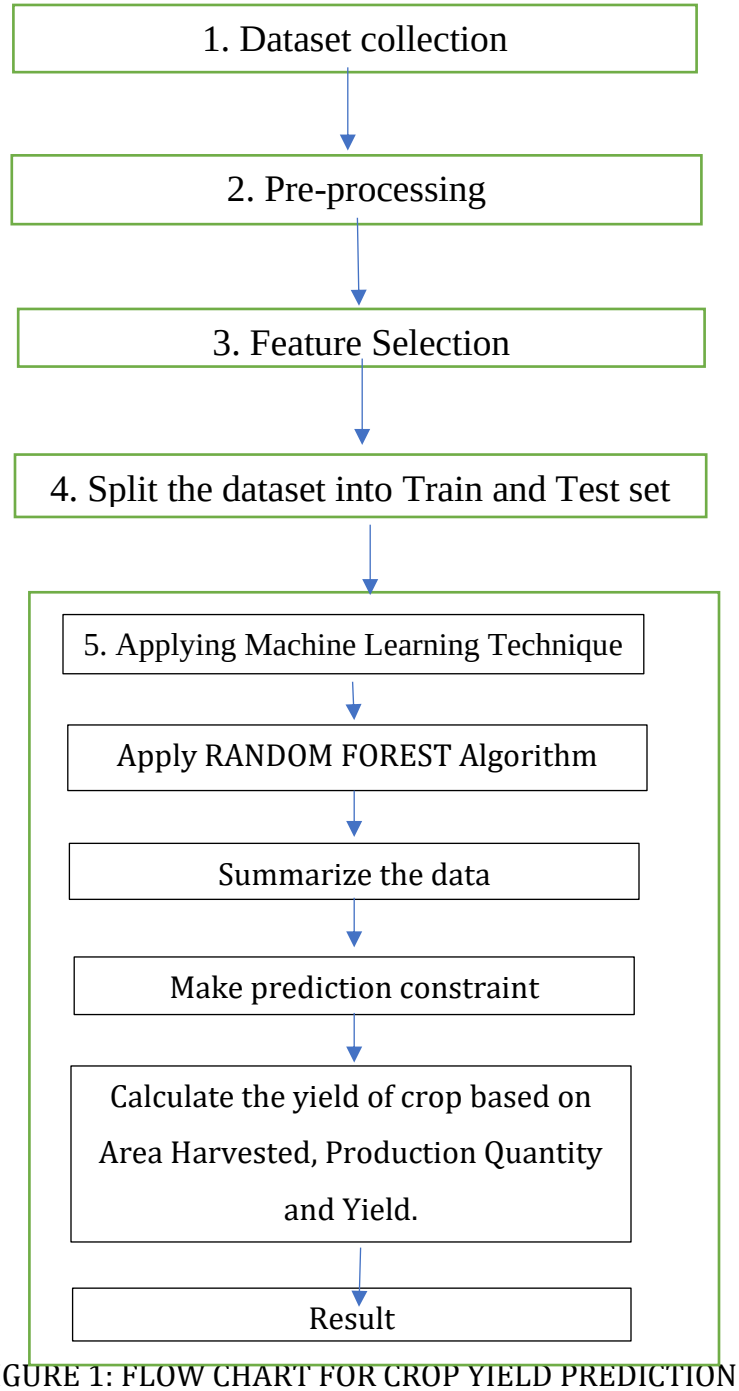


FIGURE 1: FLOW CHART FOR CROP YIELD PREDICTION

WORD COUNT:869

4 WEEKLY RESEARCH PLAN

The preliminary analysis is going to be conducted throughout the months of April and May. The intentions are to use R studio for this thesis. Therefore, the author intends on enterprise writing categories via the assistance of knowledge datacamp an internet learning platform for the aim of changing into practice in coding. Explicit attention is going to be created on classification and information visualization. The time to write this thesis is constrained of short, intensive reading is going to be necessary to confirm the task will be less discouraging because the thoughts of writing a masterpiece in fourteen weeks is intimidating.

The following chart may be a projected arrangement on the activities that may occur throughout the required weeks and dates. because it is barely a tentative schedule it will be custom-made to satisfy the wants each of scholar and supervisor. Please note that the gaps between weeks don't mean no work is conducted on the contrary. Work is occurring usually weekday to weekday like every traditional operating weekday beginning at 9 am and finishing at roughly 6 pm. Associate in nursing hour are had for lunch and frequent breaks will be taken on not get demoralized and overpowered with the amount of work. The analysis arrangement doesn't define what's to occur every week. A special approach will be taken instead the arrangement will be written little by little of a period, three days, and eventually ending with three weeks for the finishing touches. The truth of life is that schedules are created to be broken therefore what's written helpful tips to help within the direction to confirm this thesis gets written.

The Following chart shows the 14 weeks plan for this thesis. This dissertation is deliverable in eight stages:

NAME	BEGIN DATE	END DATE	DURATION
FIRST DELIVERABLE	1/6/2020	14/6/2020	2 WEEKS
Meeting with supervisor to propose thesis	1/6/2020	2/6/2020	2 DAYS
Receive clearance from Supervisor to proceed	3/6/2020	4/6/2020	2 DAYS

Reading and sourcing relevant literature	5/6/2020	14/6/2020	10 DAYS
SECOND DELIVERBLE	15/6/2020	28/6/2020	2 WEEKS
Preparation of loading the data on computer	15/6/2020	16/6/2020	2 DAYS
Initial cleaning of data	17/6/2020	18/6/2020	2 DAYS
Continue with sourcing literature	19/6/2020	22/6/2020	4 DAYS
Data and summaries literature	23/6/2020	25/6/2020	3 DAYS
In conjunction start methods	26/6/2020	28/6/2020	3 DAYS
THIRD DELIVERABLE	29/6/2020	12/7/2020	2 WEEKS
Write literature review formal draft	29/6/2020	1/7/2020	3 DAYS
Write methods section formal draft	2/7/2020	4/7/2020	3 DAYS
Submit literature review and methodologies	5/7/2020	6/7/2020	2 DAYS
Prepare rough draft for coding	7/7/2020	8/7/2020	2 DAYS
Start coding	9/7/2020	12/7/2020	4 DAYS
FOURTH DELIVERABLE	13/7/2020	26/7/2020	2 WEEKS
Make changes according to feedback	13/7/2020	19/7/2020	7 DAYS

Resubmit literature review and methodologies	20/7/2020	26/7/2020	7 DAYS
FIFTH DELIVERABLE	27/7/2020	9/8/2020	2 WEEKS
Finish coding	27/7/2020	3/8/2020	8 DAYS
Write conclusion and evaluation	4/8/2020	9/8/2020	6 DAYS
SIXTH DELIVERABLE	10/8/2020	23/8/2020	2 WEEKS
Submit work to the supervisor	10/8/2020	18/8/2020	9 DAYS
Design Poster	19/8/2020	23/8/2020	5 DAYS
SEVENTH DELIVERABLE	24/8/2020	30/8/2020	1 WEEK
Write abstract and introduction	24/8/2020	27/8/2020	4 DAYS
Get feedback and final resubmission	28/8/2020	30/8/2020	3 DAYS
EIGHTH DELIVERABLE	31/8/2020	6/9/2020	1 WEEK
Final Editing and putting together complete version	31/8/2020	3/9/2020	4 DAYS
Print, bind and submit full dissertation	4/9/2020	5/9/2020	2 DAYS
Print poster	6/9/2020	6/9/2020	1 DAY

TABLE 2: WEEKLY RESEARCH PLAN FOR 14 WEEKS

The Following Gantt chart shows the 14 weeks plan for this proposed thesis. This Gantt chart is made by using GAnTTPROJECT tool, which is an open source platform.

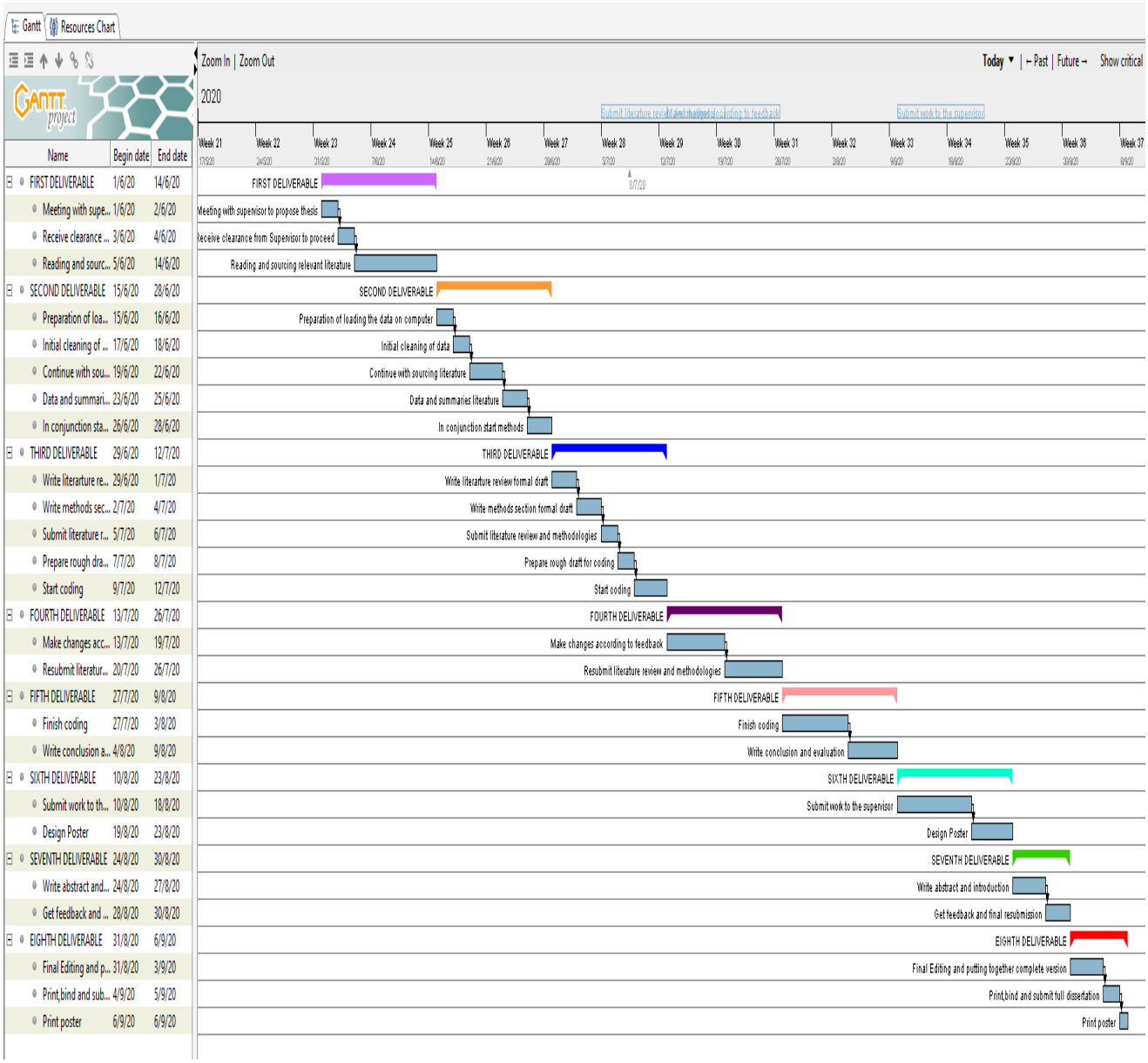


FIGURE 2: GANTT CHART FOR 14 WEEKS RESEARCH PLAN

5 REFERENCES

Citations by using IEEE style

[1] J. Ramirez-Villegas and A. Challinor, "Assessing relevant climate data for agricultural applications," *Agric. For. Meteorol.*, vol. 161, pp. 26–45, Aug. 2012, doi: 10.1016/j.agrformet.2012.03.015.

[2] J. Frausto-Solis, A. Gonzalez-Sanchez, and M. Larre, "A New Method for Optimal Cropping Pattern," in *MICAI 2009: Advances in Artificial Intelligence*, Berlin, Heidelberg, 2009, pp. 566–577, doi: 10.1007/978-3-642-05258-3_50.

[3] B. Devika and B. Ananthi, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUE TO PREDICT ANNUAL YIELD OF MAJOR CROPS," vol. 05, no. 12, p. 6, 2018.

[4] "Machine Learning in Agriculture: Applications and Techniques," *KDnuggets*. https://www.kdnuggets.com/machine-learning-in-agriculture-applications-and-techniques.html/ (accessed Apr. 17, 2020).

[5] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, p. 6, 2002.

[6] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, May 2015, pp. 138–145, doi: 10.1109/ICSTM.2015.7225403.

[7] S. Mishra, D. Mishra, and G. Santra, "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper," *Indian J. Sci. Technol.*, vol. 9, Oct. 2016, doi: 10.17485/ijst/2016/v9i38/95032.

[8] E. Manjula and S. Djodiltachoumy, "A Model for Prediction of Crop Yield," vol. 6, no. 4, p. 8, 2017.

[9] K. Kaur, "Machine Learning: Applications in Indian Agriculture," vol. 5, no. 4, p. 3.

[10] B. Ji, Y. SUN, S. YANG, and J. WAN, "Artificial neural networks for rice yield prediction in mountainous regions," *J. Agric. Sci.*, vol. 145, Jun. 2007, doi: 10.1017/S0021859606006691.

[11] N. Jain, A. Kumar, S. Garud, V. Pradhan, and P. Kulkarni, "Crop Selection Method Based on Various Environmental Factors Using Machine Learning," vol. 04, no. 02, p. 4.

[12] D. Ramesh, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES," *Int. J. Res. Eng. Technol.*, vol. 04, no. 01, pp. 470–473, Accessed: Apr. 17, 2020. [Online]. Available: https://www.academia.edu/21226449/ANALYSIS_OF_CROP_YIELD_PREDICTION_USING_DATA_MINING_TECHNIQUES.

[13] A. K. Kushwaha and M. Tech, "Crop yield prediction using Agro Algorithm in Hadoop," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, p. 4, 2015.

[14] Vivekanand College, A. A. Raorane, R. V. Kulkarni, and Chh. Shahu Institute of business Education and Research Centre, "APPLICATION OF DATAMINING TOOL TO CROP MANAGEMENT SYSTEM," *Russ. J. Agric. Socio-Econ. Sci.*, vol. 37, no. 1, pp. 3–16, Jan. 2015, doi: 10.18551/rjoas.2015-01.01.

[15] A. Savla and P. D. Himtanaya Bhadada, "APPLICATION OF MACHINE LEARNING TECHNIQUES FOR YIELD PREDICTION ON DELINEATED ZONES IN PRECISION AGRICULTURE," *Int. J. New Comput. Archit. Their Appl. IJNCAA*, vol. 5, no. 2, pp. 48–52, 2015, Accessed: Apr. 17, 2020. [Online]. Available: http://sdiwc.net/digital-library/application-of-machine-learning-techniques-for-yield-prediction-on-delineated-zones-in-precision-agriculture.

[16] X. Morin and W. Thuiller, "Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change," *Ecology*, vol. 90, no. 5, pp. 1301–1313, Accessed: Apr. 18, 2020. [Online]. Available: https://www.academia.edu/13787528/Comparing_niche-_and_process-based_models_to_reduce_prediction_uncertainty_in_species_range_shifts_under_climate_change.

[17] A. Prasad, L. Iverson, and A. Liaw, "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction," *Ecosystems*, vol. 9, pp. 181–199, Mar. 2006, doi: 10.1007/s10021-005-0054-1.

[18] M. Marmion, M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller, "Evaluation of consensus methods in predictive species distribution modelling," *Divers. Distrib.*, vol. 15, no. 1, pp. 59–69, 2009, doi: 10.1111/j.1472-4642.2008.00491.x.

[19] D. R. Cutler *et al.*, "RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY," *Ecology*, vol. 88, no. 11, pp. 2783–2792, Nov. 2007, doi: 10.1890/07-0539.1.

[20] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, Jan. 2006, doi: 10.1186/1471-2105-7-3.

[21] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests," *BMC Genet.*, vol. 5, no. 1, p. 32, Dec. 2004, doi: 10.1186/1471-2156-5-32.

[22] S. Fukuda, W. Spreer, E. Yasunaga, K. Yuge, V. Sardsud, and J. Müller, "Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes," *Agric. Water Manag.*, vol. 116, pp. 142–150, Jan. 2013, doi: 10.1016/j.agwat.2012.07.003.

[23] O. Mutanga, E. Adam, and M. A. Cho, "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 18, pp. 399–406, Aug. 2012, doi: 10.1016/j.jag.2012.03.012.

[24] S. Vincenzi *et al.*, "Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy," *Ecol. Model.*, vol. 222, no. 8, pp. 1471–1478, Apr. 2011, doi: 10.1016/j.ecolmodel.2011.02.007.

[25] U. Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *Am. Stat.*, vol. 63, no. 4, pp. 308–319, Nov. 2009, doi: 10.1198/tast.2009.08199.

[26] R. Berk, "Statistical Learning From a Regression Perspective," 2008, p. 358.

[27] "FAOSTAT." http://www.fao.org/faostat/en/#data/QC (accessed Apr. 21, 2020).

[28] "randomForest function | R Documentation." https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest (accessed Apr. 21, 2020).