**CROP YIELD PREDICTION USING MACHINE LEARNING**

Table of Contents

Acknowledgments 2

List of Figures 5

Abstract 7

1 Introduction 8

2 STEEPLE Analysis 9

2.1 Societal Factor 9

2.2 Technological Factor 9

2.3 Environmental Factor 9

2.4 Ethical Factor 9

2.5 Political Factor 10

2.6 Legal Factor 10

2.7 Economic Factor 10

3 Literature Review 10

3.1 Supervised Learning 11

3.1.1 Decision Tree 11

3.1.2 Naïve Bayes 13

3.1.3 K -Nearest Neighbour (KNN) 13

3.1.4 Support Vector Machine (SVM) 14

3.2 Semi -Supervised Learning 14

3.3 Unsupervised Learning 14

3.3.1 Principal Component Analysis (PCA) 15

3.3.2 Clustering with K -Means 16

4 Problem Setting 18

5 Methodology 18

5.1 Dataset 18

5.2 Method Used 20

5.3 Techniques Use d 21

5.3.1 Orange Data Mining 21

5.3.2 RStudio 21

5.3.3 Excel 22

5.4 K -means Clustering 22

6 References 32

**List of Figures**

FIGURE 1 SUPERVISED LEARNING TECHNIQUE [1]...................................................11

FIGURE 2 NAÏVE BAYES THEOREM .............................................................................13

FIGURE 3 EUCLIDEAN DISTANCE [10] .........................................................................16

FIGURE 4 ASSIGNMENT [10]...........................................................................................16

FIGURE 5 FINDING THE NEW CENTROID [10].............................................................17

FIGURE 6 OUTLIERS TEST..............................................................................................20

**List of Tables**

TABLE 1 CROP YIELD PREDICTION DATASET............................................................19

TABLE 2 PERCENTAGE OF GOOD SAMPLES ...............................................................20

**ABSTRACT**

As Morocco is going green, it is tackling many difficulties in agriculture like climate change, drought, desertlike and saltiness. These issues acquire many moves to save them. And for that, Morocco introduced the production of a crop in the early 2000s that have the characteristics needed for these issues. Thus, the main purpose of this combined project is to provide farmers with access to suitable crop management methods that will lead them to high crop production and strong value chains between Morocco and the United Kingdom. This research was mainly done by analysing different machine learning techniques used in agriculture that help to minimize the wastes of smallholder farmers and maximize their production yields. Then, it intends to provide a prediction of any crop yield from previous data. And it is mainly done by applying K-means clustering on agricultural features. It aims, therefore, to predict crop yields for sustainable agriculture, and it was supposed to be applied for Quinoa's crop yields in Rhamna region. But, due to what is happening in the world these days with COVID-19, I could not get the relevant data about Quinoa's production in the selected region. So, I used some data from Kaggle, a website that has powerful resources for data science.

**Chapter 1**

**INTRODUCTION**

Nowadays, quinoa is becoming more and more popular not only because of its health benefits but also for its high revenue-generating. At the beginning of its popularity, only a few countries were producing it such as Bolivia, Peru, and Chile. But now, it is getting expanded in the world. Morocco was been introduced to Quinoa in the 2000s. But its production has been constrained by the shortage of access to well-adapted and high-yielding cultivars, inappropriate crop management practices, weak value chains, and limited market demand [14]. The aimed project was supposed to be executed in Rhamna Region, where a plethoric number of smallholder farmers live under the edge of poverty and where a quinoa value chain already exists but is overly controlled by several circumstances. The aim of this capstone project is not only to bring an innovative and ethical idea to be used in agriculture needs but also to create jobs in Morocco through digital and technological solutions. It is proposed to better food security and to improve the income of smallholder farmers in Rhamna Province. Furthermore, this project is a joint work between two SBA students and me. My two BA teammates will work on the business part and I will tackle the technical one. The technical part of this project consists of implementing and using machine learning techniques to predict crop yields; this is mainly to optimize the production of the crop and to better the income of the smallholder farmers in Morocco. And to do so, three main tools will be used to accurately reach the objectives of this project. Orange Data Mining (software) is a great starting point especially for beginners in the field as it is open-source and it helps to easily apply machine learning and data visualization techniques to see which works best in any case scenario. For me, I will use it just to see if there are any outliers in the chosen dataset. As for the model and language to be used, after thorough research, the programming language to be used for implementation will be RStudio, this is due to its exemplary support for data wrangling and the availability of pre-built designs and open-source particularly useful libraries that can help with my goal and maybe even enhance results. And finally, I will use Excel to identify the differences between executing RStudio on my Training and Testing sets. This will show me the accuracy of the results as I cannot directly measure the accuracy of the applied K-means clustering.

**Chapter 2**

**STEEPLE ANALYSIS**

**2.1 Societal Factor**:This work is not only beneficial to corporations but also smallholder farmers. This prediction model will support the social development of rural areas in Morocco, especially the smallholder farmers who are not precise with their crop yields. And then, it can be applied in other countries, in which agriculture is the most significant sector, by minimizing the wastes of the production and maximizing the income of smallholder farmers.

**2.2 Technological Factor:** Machine Learning boosted greatly in solving problems in all the areas, especially the agricultural sector. So, the prediction of crop yields based on data mining techniques will be developed in a way that permits continuing it with other technologies in the future. This determines the development of all related technology which could signify a significant improvement of all data mining techniques used for prediction. And, thus, in this project, I will implement K-means clustering to predict high production crop yields using RStudio and Orange Data Mining software.

**2.3 Environmental Factor:** This model will provide a significant narrowing of the negative effects of climate variability and extremes under present-day climate requirements. Because it will help the farmers to know the precise productive features that can provide them with high crop yields. This will also allow farmers to produce a high yield while minimizing the wastes by diminishing the used resources that can harm our environment. Ant this will lead to reducing the pollution caused by farmers.

**2.4 Ethical Factor:** The prediction model using K-means clustering was mainly based on including data from some smallholder farmers. But, with what is happening these days with COVID-19, I could not reach the data needed so I used crop yield prediction data from Kaggle. Indeed, there is no 10 unethical use of the chosen data. And this work does not damage any party, but it benefits the agricultural sector of our country.

**2.5 Political Factor**: The Moroccan government is highly encouraging farmers to produce more in the agricultural sector. However, there is no clear correlation between politics and forecasting crop yields.

**2.6 Legal Factor:** The prediction model using K-means clustering does not depend on any illegal data. So, I will not be using any illegal data from.

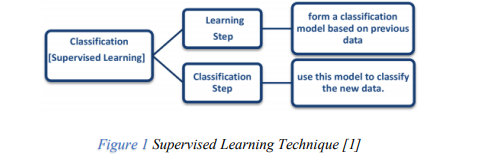
**2.7 Economic Factor:** This work has a high correlation between the economical aspect and predicting crop yields. The produced model using K-means clustering intends to commit to the improvement of the agricultural sector. It proposes to improve the crop yield and decrease the costs required for growing the crop yields because the smallholder farmers will know what exactly they need to produce effective high crop yields. Therefore, this work will help smallholder farmers and improve their economic situation, which will be beneficial to our country.

**Chapter 3**

**LITERATURE REVIEW**

Machine learning is a process of making computers to make something intelligent based on some previous bunch of data. So, it could be developing them on how to perform some difficult duties, something that we could not just duplicate all the precepts for like using a motorized vehicle without a chauffeur. Moreover, it could be applied to mine a big dataset to detect models and correlations that we could not necessarily be able to perceive as human beings. And I noticed that people may have learned of the word ML, but they do not recognize its real meaning. Nowadays, it is a helpful point to understand about the machine learning' concept. We cannot measure how we decently interact with ML algorithms all the time, whenever we use our laptops, we binge-watch, or we surf on the net; they predict and command us a lot of what occurs in our days. A simple example of machine learning 11 applications is Netflix or Amazon; when they dictate showing you what other stuff you could like to watch or purchase. Another example is when your email can filter spam on our inboxes. Machine learning is a subset of Artificial Intelligence (AI). And we can define it as merging of statistical and computer science tools to explore, analyze and visualize data. In Machine Learning, we have three different approaches: supervised learning, semi-supervised learning and unsupervised learning.

**3.1 Supervised Learning:** Supervised learning or classification as it is known in the world of machine learning is split into a two-step process[1] . It is a method that builds forecasting models from data. The models are acquired by recursively dividing the data space and providing a simple prediction model within each division. Figure 1 notes that the first part is the learning step; this is where we work on a model that was built from the previous data set. The second part is the classification part where we deploy the model to classify the new data if its correctness is appropriate [1]. Next are the common classification methods.



**3.1.1 Decision Tree :**

A decision tree is one of the most popular supervised learning methods. It is used to make decisions precisely and accurately. As it is named, it applies a tree-like model of decisions. And it can deal with both categorical and numerical data. There are many algorithms used in the decision tree, but here, we will focus on J48 and Simple Cart.

**3.1.2 Naïve Bayes :**

Naïve Bayes is another algorithm used for statistical classification based on the Bayes rule. It implements a probabilistic forecast. The attributes are conditional independence between classes. And the algorithm tends to measure these conditional probabilities of the samples given.

**3.1.3 K-Nearest Neighbour (KNN) :**

The KNN is one of many classification algorithms used in machine learning. It is a classifier algorithm where the learning is based on the similarities of data from others. This tool is nonparametric as it does not make any assumption on data distribution. And it does not really learn any model and make generalization of the data. So, this algorithm focuses only on classifying objects based on similar characteristics.

**3.1.4 Support Vector Machine (SVM) :**

SVM is a description of the training data set like points in space divided into categories by a clear gap that is as scattered as possible. The new examples that we are looking for are then mapped into that same space and forecasted to refer to a category based on which side of the gap they fall. This can simply mean that this technique relies on specified input data to determine a function that produces a relevant output when given new unlabeled data. The principal role of SVM is to increase the space between the two classes to decrease the error when the given data are classified. This tool is powerful when it comes to high dimensional spaces and it uses a subset of sampling points in the decision function, so it is also memory efficient [5].

**3.2 Semi-Supervised Learning:**

The framework in reinforcement learning is surprisingly similar to the normal framework of supervised learning. We have an input frame, we run it through some neural network model, and the network produces an output action. The only difference here is that now we do not know the target label. We do not know in any situation whether we should have gone, because we do not have any dataset to train on. Reinforcement learning is a learning standard which deals with the study of how computers and natural systems such as human beings acquire knowledge in the presence of both labelled and unlabeled data [8]. This concept was mainly made to rectify the drawbacks of supervised and unsupervised learning. Normally, this mixture will include a slight number of labelled data and a big number of unlabeled data. The primary method required is that the programmer will cluster related data using an unsupervised learning algorithm and then use the existing labelled data to label the rest of the unlabeled data [9].

**Chapter 4**

**Problem Setting**

In Morocco, the agricultural sector is showing a flourishing growth and accomplishments. It was published by the World Bank collection of development indicators that 37.86 % of Moroccans are working in the agricultural sector in 2019, which shows that more than the third of our population is living on the field of agriculture [11]. However, we cannot deny the challenges and issues this sector is facing as every country in the world. To do down some of these issues, I chose to work on this project to better the crop yield production by using some previous specific data to predict the crop yields. The crop yield is the main component to decide whether the farmer is doing well or not. And by controlling it, we can optimize the production of this crop while we are minimizing its wastes. This project was at first mainly done for Quinoa's production in Rhamna region. But, due to what is happening in the world with COVID-19, I made it more general and can be applied to any kind of crops while we have a sufficient amount of required data. By applying K-means clustering on the data that I am going to discuss later on in this work, we can make decisions of the outputs whether the crop yield will be good or not. This technique will help smallholder farmers to minimize the losses and maximize their crop production

**Chapter 5**

**Methodology**

**5.1 Dataset:**

In any ML analysis, you need data. And any model can only be powerful if you feed it with the right data. The on-target data should have the precise features and the right outcomes because it will affect the relevance and the usability of the model as well as the findings. The data applied for my work was obtained from Kaggle, the world's largest data science community with powerful tools and resources. It was made for crop yield prediction. The data has 11 features and one target. These features are:

1. Moisture: the value of moisture in the soil can enormously affect the production of the crop

2. Rainfall: it is measured in millimeters per hour in the sample

3. Average Humidity: it is measured in grams of water vapor per cubic meter volume of air

4. Mean Temperature: it measured in the Celsius scale

5. Maximum Temperature: it measured in the Celsius scale

6. Minimum Temperature: it measured in the Celsius scale

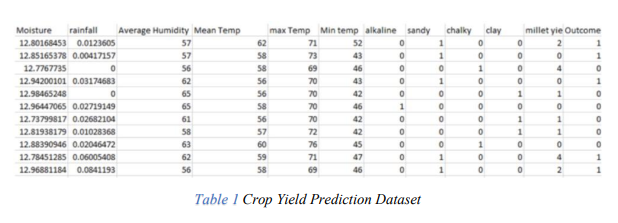
7. Alkaline: it specifies if the soil is alkaline. Alkaline soil is when the pH is above 7.

8. Sandy: it specifies if the soil is sandy

9. Chalky: it specifies if the soil is chalky

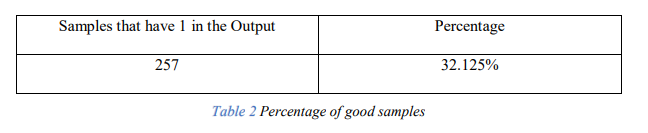
10. Clay: it specifies if the soil is clay

11. Millet Yield: it specifies in ton per hectare And the target is the Output. When the output is one, it means that the sample has a high production of the crop yield, and zero is the opposite.



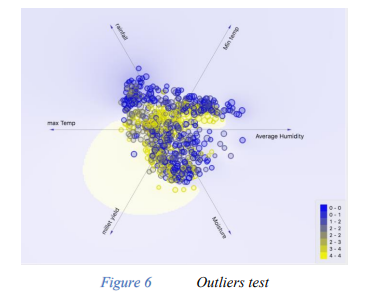
Above is a screenshot of my dataset. My output is the Outcome feature; it predicts the environment in which the outcome is productive or not.

Below, a table that shows the percentage of the samples that have 1 in the outcome. From this result, I have to take into consideration this percentage in my K-means clustering analysis because the results will surely have more zeros' values.



**5.2 Method Used:**

The main point of this study is to predict from previous crop yields data the outcome of a sample using the last technique discussed in the unsupervised learning section. Therefore, I will apply K-means clustering to my data set using RStudio and Orange Data mining software. So, the methodology that I used has two phases: Training and Test phases. In the first part, I looked for the data, I preprocessed it and I applied linear projection using PCA to detect if I have any outliers using Orange Data Mining tool. Linear projection is an algebra notion that benefits to differentiate between the outliers and the inliers in a data set. This can help to determine the right features of the model used in the prediction. The results shown in the figure below revealed that I don’t have any outliers.



**5.3 Techniques Used:**

In this work, I will implement the K-means clustering for a predictive analysis in Orange Data Mining software, RStudio and Microsoft Excel.

**5.3.1 Orange Data Mining :**

Orange Data Mining is, first and foremost, free and open-source tool. It helps to easily apply machine learning, data mining and data visualization techniques to analyze data. I will be mainly using it for the preprocessing part of my data set. By using it, I can determine whether my data set has outliers or not. This step will help me to get rid of the inaccurate samples from my data set and avoid significant issued during my clustering analysis.

**5.3.2 RStudio :**

RStudio is an integrated development environment (IDE) for R [15]. R is free and open source. It is a language that is not difficult to use and relatively intuitive. Another reason for using R is because there are people from all over the world writing packages; these packages contain functions and data sets that you can install and use for free. Moreover, R has an extraordinary data visualization as well as graphics abilities. Before the implementation, I had to call some libraries and install some packages to run the K-mean clustering.

These packages are:

• library(tidyverse): this package is used to manipulate the data.

• library(cluster): this package is used to implement the clustering algorithms.

• library(factoextra): this package is used to extract and visualize the data.

• library(gridExtra): this package is used to arrange multiple plots.

**5.3.3 Excel :**

Microsoft Excel is one of the most significant tools to analyze data. I will mainly use it to classify and analyze the differences between the results of RStudio K-means clustering in the training and testing sets. As K-means clustering is part of the unsupervised learning, I cannot directly measure the accuracy of my results. But, I used two methods in RStudio to identify the optimal number of clusters to assure the right results.

**Chapter 6**

**References**

[1] Patel, Hetal, and Dharmendra Patel. “A Comparative Study on Various Data Mining Algorithms with Special Reference to Crop Yield Prediction.” Indian Journal of Science and Technology, vol. 9, no. 22, Dec. 2016, doi:10.17485/ijst/2016/v9i22/92713.

[2] Kaur, Gaganjot, and Amit Chhabra. “Improved J48 Classification Algorithm for the Prediction of Diabetes.” International Journal of Computer Applications, vol. 98, no. 22, 2014, pp. 13–17., doi:10.5120/17314-7433.

[3] Ijcsis, J. “Naïve Bayes and J48 Classification Algorithms on Swahili Tweets: Perfomance Evaluation.” Academia.edu - Share Research\_Tweets\_Perfomance\_Evaluation .

[4] Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. pdfs.semanticscholar.org/26d6/73f140807942313545489b38241c1f0401d0.pdf.

[5] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, International Journal of Computer Science and Technology IJCST Vol. 2, Issue 1, March 2011.

[6] Jolliffe, I. T., Cadima, J., Ian T. Jolliffe Ian T. Jolliffe, Jolliffe, I. T., Cadima, J., Jorge Cadima Secção de Matemática, … Ian T. Jolliffe College of Engineering. (2016, April 13). Principal component analysis: a review and recent developments. Retrieved from <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>

[7] Dabbura, I. (2019, September 3). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Retrieved from https://towardsdatascience.com/kmeans-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[8] Jothi, V., & M., L. (2014, February 19). A Survey on Semi-Supervised Learning Techniques. Retrieved from <https://arxiv.org/abs/1402.4645>

[9] AlindGuptaCheck. (2019, May 29). ML: Semi-Supervised Learning. Retrieved from <https://www.geeksforgeeks.org/ml-semi-supervised-learning/>

[10] Krishnan, M. (2018, July 7). Mathematics behind K-Mean Clustering algorithm. Retrieved from https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/