Introduction

Machine learning (ML) enables a system to scrutinize data and deduce knowledge. It goes beyond simply learning or extracting knowledge, to utilizing and improving knowledge over time and with experience. In essence, the goal of ML is to identify and exploit hidden patterns in “training” data. The patterns learnt are used to analyze unknown data, such that it can be grouped together or mapped to the known groups.This instigates a shift in the traditional programming paradigm, where programs are written to automate tasks. ML creates the program (i.e. model) that fits the data. Recently, ML is enjoying renewed interest. EarlyML techniques were rigid and incapable of tolerating any variations from the training data. Recent advances in ML have made these techniques flexible and resilient in their applicability to various real world

scenarios, ranging from extraordinary to mundane. For instance, ML in health care has greatly improved the areas of medical imaging and computer-aided diagnosis.Ordinarily, we often use technological tools that are founded upon ML. For example, search engines extensively use ML for non-trivial tasks, such as query suggestions, spell correction, web indexing and page ranking. Evidently, as we look forward to automating more aspects of our lives, ranging from home automation to autonomous vehicles, ML techniques will become an increasingly important facet in various systems that aid in decision making, analysis, and automation.part from the advances in ML techniques, various other factors contribute to its revival. Most importantly,the success of ML techniques relies heavily on data .Undoubtedly, there is a colossal amount of data in todays’ networks, which is bound to grow further with emerging networks, such as the Internet of Things (IoT) and its billions of connected devices . This encourages the application of ML that not only identifies hidden and unexpected patterns, but can also be applied to learn and understand the processes that generate the data.

Agriculture, since its invention and inception, has been the prime and pre-eminent activity of every culture and civilization throughout the history of mankind. It is not only an enormous aspect of the growing economy, but it’s essential for us to survive. It’s also a crucial sector for the Bangladesh economy and also human future. It also contributes an outsized portion of employment. Because time passes the requirement for production has increased exponentially. So as to produce in mass quantities people are using technology in an exceedingly wrong way. New sorts of hybrid varieties are produced day by day. Bangladesh has been surviving on agriculture from the very beginning. Agriculture has been playing the main part of our Bangladeshi economy. As we are in the age of modern science and technology, so for developing in this field we need to apply the best possible option and it will automatically develop the Bangladeshi economy. If we are able to apply our best work to improve the production rate of our crops ,it will help the government to take the proper steps to handle food risk management. The farmers need to be engaged by learning the methods and utilize the techniques of smart farming. Every computer engineer has to support the farmers with the new technology to amplify the productivity of a crop. Also, one more issue turns up

whether the farmers believe the benefit of technology, probably ‘no’. The only way to

reach the farmers is by the proven results, already numerous techniques have been

projected for this view and which has provided good results.

However, these varieties don’t provide the essential contents as naturally produced crops. These unnatural techniques spoil the soil. It all ends up in further environmental harm. The core emphasis would be on precision agriculture, where quality is ensured over undesirable environmental factors. So as to perform accurate prediction and stand on the inconsistent trends various machine learning classifiers like, Naïve Bayes, Random Forest, Support Vector machine etc. are applied to urge a pattern.

**B: Data Preprocessing and analysis**

Data preprocessing is the second step and it contains two steps. Original dataset can contain lots of missing values so initially all these should be removed. Missing values are denoted by a *NaN* in the dataset and their presence can deteriorate the value of the entire data and it can reduce the performance.

So we took the dataset from Koena website, where it holds agricultural records for almost all the countries that include features like area year-wise harvested, yield, production and their respective unit values. The dataset had flaws like missing entire rows of data, data that are not relevant to our research. We modified the dataset according to our requirement and processed it manually.

Dataset preprocessing

How to handle nominal categorical data, using one hot encoding. We had to convert categorical data to numerical data while working on this dataset, the reason being, a handful of algorithms don't accept strings. Categorical data is of two types, one ***nominal***, they don’t have an intrinsic order and the other one is ***ordinal***, where the data follows an intrinsic order. With ordinal data we use ordinal encoding. But our data required a nominal encoding where we used **One Hot Encoding.** As our dataset contains the data for all crops in Bangladesh, there are lots of categories of data without an intrinsic order. So to convert them into numbers, we cannot use an ordinal encoding as our models might think one category is greater quantitatively speaking. So the solution to this is, we create a separate column for each category. We convert strings to vectors of each category and that is how One Hot Encoding is done. No matter how many categories there are, we’ll have to make those columns.

Another concept surfaces which is a dummy variable trap, generally when we do one hot encoding, we remove one of the columns from the newly created categorical columns.

If there are ***n*** number of columns, then we will keep ***n-1*** number of columns. This brings up concepts like ***multicollinearity***, meaning the input columns have a mathematical relationship, if yes, they are dependent on each other and it should be avoided. The columns should be independent, and y is a dependent column as it depends on the independent columns. We can not let the columns have a mathematical relationship. The newly created columns are called dummy variables which raises the problem of multicollinearity, that is why it is called the Dummy variable trap. The solution to that is keep ***n-1*** columns.

We have used One Hot Encoding using Sklearn, and removed multicollinearity by removing the categories from the column with drop first and setting the data type as an integer.