DATA Technologies in Agriculture

Chapter 2: Data Technologies in Agriculture

Introduction

Since ancient times, farming has been an information-based profession. The farmers were growing different types of plants. They used to keep the crop with the most productivity and the best taste. Sharing the successful agricultural practices has played a significant role in the spread of agriculture and covering the needs of millions of people over the years. For example, the cultivation date was (and still is, to some extent) dependent on trial and error. Farmers planted on a specific date; the crop grew well initially but was attacked by frost a month before the harvest, which destroyed the crop. Farmers learn and decide to change the planting date to be two months before the old date to avoid frost. They see that the growth of the crop is weak compared to what it was previously, but the growing cycle is completed successfully till the harvest. So, they decide to add two weeks to the second date in the following year until they reach the best date for planting. The same applies to the quantities of irrigation water, tillage practices, and all other farming operations.

Farmers from other regions took some seeds to plant in their fields. They took the seeds, but they also took instructions or 'information' about sowing dates, cultivation methods, irrigation quantities, etc. However, they had to tolerate these practices to suit their environment. This information is a set of data that integrate to form a logical meaning.

The cultivation instructions in the past were inaccurate and straightforward. The farmers used to say, for example, "The land must be plowed well," or "Open the irrigation gates for two hours." This information has a broad meaning and is not generalizable, as the concept of good plowing for a farmer is not the same for others. Some farmers can apply their 'good' tillage to such an extent that it causes problems in irrigation or any other problems. Similarly, irrigation for two hours may be suitable for one type of soil and be too few or too much for other soils.

With data acquisition and transmission technologies, the concepts have changed to be more specific, like to till the soil to a specific depth to reach a certain bulk density. Likewise, irrigation is started when the soil moisture reaches a specific rate and ends when reaching the field capacity. The advances in sensors' technologies make data more accurate, specific and customized for field conditions.

Agricultural operations are now conducted in a stream of data. Every modern farm now has its weather data, soil data, and crop-related data (like water requirements) in addition to regional data about insects, pests, diseases. Financially, farmers have data about potential markets, including transport expenses and possible selling prices. These data help the farmers to plan for crop and varieties selection and the harvest time to maximize their gain. Nevertheless, the advanced analysis of data and artificial intelligence techniques made it possible to predict crop productivity and selling prices, reduce the use of pesticides and chemicals to a minimum, and many other benefits discussed in this chapter. Agriculture has become a data-driven industry.

Data types

Data has many formats; some are simple like numeric and string formats, while some are complex, like images, audio, and video. Regardless of the format, all kinds of data are stored digitally in binary form, i.e., ones and zeros. The more the data complexity, the larger its binary size. The binary size affects data transmission speed and the space required for data storage. The simplest data formats are the numeric format, then simple texts (like names and addresses), then complex texts (like reports and emails), then

images and audio (varies depending on the quality), and finally, the video format (which is a combined image series with audio).

Data acquisition and transmission

Data is generated by sensors; for example, a soil-moisture sensor measures a physical phenomenon that changes with water level (electric conductivity, capacitance, or tension). It converts the phenomena to numbers that can be transferred to be stored in some digital medium. To use this data, field calibration must be performed to convert the raw data to data with physical meaning. The same concept applies to drones images captured through imagery sensors attached to the drone, then transferred wirelessly to a cloud server or directly to a local PC. Then the RAW images are analyzed to extract the desired features.

As we see, all data types have similar life cycles, from sensors to computers. Data transmission is performed via several means, either wirelessly or directly (via USB or any wired connection.) through the transmission operation. Some forms of data are secured to ensure data safety and accuracy. More information and details about data transmission are in the IoT chapter. The last destination of data is the storage, which is performed either using conventional databases (local or cloud) or via distributed ledger databases (details and benefits of this technology are in the DLT chapter)

A schematic representation of the data lifecycle is shown in Figure 12. Data is born by different sensors (including drone pictures) or by taking text or voice notes. Next, data is transmitted through different means to the cloud or local storage. After that, data analysts clean the data by merging, filtering, removing duplicates, and filling gaps. Then, the analysis stage started by finding insights, performing predictions, and discovering relationships. The findings are then shared with stakeholders through reports and presentations to act. Finally, the data is either being archived for further analyses or being deleted safely.

Data V's

As the sensors' technology advances, sensors become cheaper and smaller, becoming everywhere in modern farm facilities. The increasing number of sensors yields a massive amount of data to be stored and analyzed. Conventionally, analyzing data in a computer was limited to the system's memory. If the data volume is larger than the system's memory, it needs special handling tools. This data is called Big Data technology. Big data is generally defined as data with three main features, Volume, Velocity, and/or Variety, as defined by Gartner Inc. in 1997. "Volume" stands for the cumulative data amount and data stream. "Velocity" reflects how quickly the data is generated and the speed needed for it to be analyzed and be ready to use. "Variety" reflects that different data types form the complete picture. For example, to analyze crop water needs, we might need to analyze numeric data from the meteorological station and the soil water sensors, along with satellite images, drone images. Different data types represent the variety that forms the complete picture.

Figure 12. Data lifecycle (by the author)

As we notice, the three main features of Big Data; Volume, Velocity, and Variety, all start with V, which is called the 3 V's of data. In 2013, IBM added another V, which is Veracity, which reflects the level of trust of data, i.e., its trustfulness and completeness. This is mainly raised for sensitive fields that require the picture to be complete, like medical fields. If we have untrusted data, it may lead to disastrous decisions. Later, different researchers added more V's as characteristics of big data. For instance, "Value" stands for how the data is worth, "Variability" refers to the nature of data that changes despite

all other variables remaining constant. For example, if we take soil data for the same field, it will be variable every day because it is a dynamic media. The same applies to weather data.

Additionally, we have "Variability" that should not be confused with variety, as variety is about different data types in the same dataset, while variability is within the same feature. Microsoft also added the 6th V: "Visibility", which stands for how the data form the complete picture to make the right decision. Other V's are "Validity," which shows the accuracy of data, "Volatility," which shows the age of which the data can be used before it expires, and finally, "Visualization" which reflects how complex is it to show meaningful visualizations from the current data volume (99–102). A diagram showing the different V's of Big Data is shown in Figure 13.

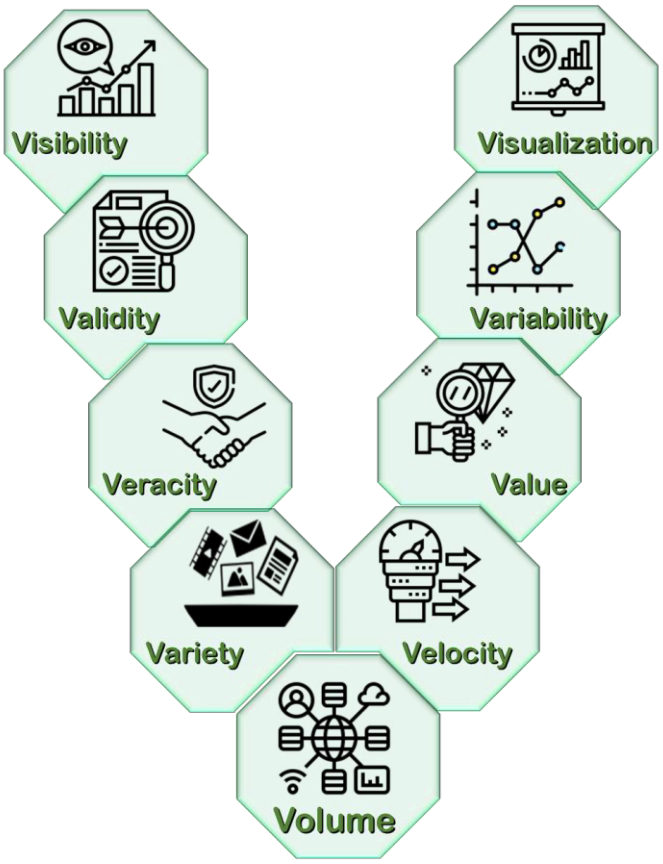

Figure 13. The nine V's of Big Data. (by the author)

Data storage and security

With the vast increase in data volume generated by smart agriculture, there is a need for reliable data storage solutions. Data storage is essential at every stage of data analytics. In many cases, we need to store the same data at every stage; raw, after cleaning, after merging, and after modeling and making predictions. Data storage can apply to physical and electronic storage. Physical storage includes storing paper copies and so on. However, in this report, we only mean digital data storage.

Digital data storage is either performed online or offline. Online storage depends on networks and remote storage solutions, while offline storage depends on local or attached devices. Offline storage solutions are always cheaper than network devices. Data can be retrievable anytime, even without a

network connection. However, sharing stored data requires much effort and physical presence at the sharing location. The offline storage devices include hard disk drives (HDD), solid-state drives (SSD), Optical disk drives (CDs / DVDs / Blu-ray), flash drives (Thumb drives).

On the other hand, the online storage solutions allow easier data sharing, but they are more expensive and depend on reliable network connections. Like offline devices, online storage has many types. The best type of network storage is cloud storage, where the backup servers are managed and secured by a specific company that rents its full service for a monthly or annual fee. There is also hybrid cloud storage which has an intermediate layer of storage before uploading to the cloud company, the benefit of it is that it reduces the costs of frequent backups and makes the data easier to reach the intermediate stage. There are many other network storage types like flash drive arrays and hybrid flash arrays.

Data security protects data from being lost, corrupted, or accessed by unauthorized personnel. Data security practices start from the moment of generation to the moment of safe disposal through the stages of transmission, analysis, and storage. Security practices include controlling privileges of users' access, creating reliable backups, data encryption and hashing, and the safe erasure of sensitive data. Although agricultural data is less sensitive than financial data, it is also subject to threats like man-in-the-middle, spyware, denial of service, and many other threats. Data security in agriculture is becoming more critical as the agriculture industry relies on electronic transactions for e-commerce, blockchain insurance, and the food supply chain.

Data analysis and analytics

The value of data is gained due to the insights identified after analyzing it. Raw data needs to be labeled, cleaned, and merged before the analysis starts. The process of analyzing data to see its current perspective is called data analysis while analyzing data to find correlations and trends is called data analytics. Thus, data analysis processes are a subset of data analytics procedures.

Data analytics involves four major types: descriptive, diagnostics, predictive, and prescriptive. The descriptive analytics answers the question of "What happened?" while the diagnostic analytics answers "Why has that happened?". Both questions are part of the data analysis subset. The other two types are about future actions; the predictive analysis answers the question "What would happen in the future if we followed the past and current trends?" on the other hand, the prescriptive analytics answers the question "What should we change to reach a specific target in the future?"

For example, when we look at the crop production data, we notice that the yield of a crop was increasing, then it started to decrease three years ago. This is a result of descriptive analysis, the "what happened?" Question. We took a deeper look at our data, and then we noticed that we had changed the types and amounts of fertilizers three years ago. We concluded that this is "why" the yield reduction happened, which is a diagnostic analysis. We started to make scenarios of what will the yield and profits be in the coming years if we continued to use the same settings and if we changed the settings. This is the predictive analysis. Finally, we need to reach a specific yield amount, so we perform some modeling by changing fertilizers, crop varieties, irrigation methods to see which setting can lead us to our target, the prescriptive analysis.

Data science and machine learning

While data analysis answers questions like "what happened" and "why", data analytics finds correlations between variables and makes regression-based trends. Data science makes more advanced models to

get answers to more complicated questions. Data science models include clustering, pattern recognition, advanced regression models, and neural network models.

Data science models are so helpful in the agriculture field. Clustering models are used to automatically sort crops by size, color, and quality. It is also used to recognize animals' features and determine their health conditions. Pattern recognition models are used for crops' disease detection and weed detection. Regression models are used in yield prediction, to identify soil fertility, to measure sustainability, to predict energy consumption of farm processes, and many other applications (103–108). Figure 14 shows some sample data science applications in agriculture from planting to post-harvest.

Figure 14. Some data science applications in agriculture

[Illustration by Zephyr Peacock, From this web page]

Machine learning is the science that uses data science and programming to enable the computerized machine to solve specific problems without being explicitly programmed to do so. Agriculture has wide applications of machine learning, like vegetables picking by drones, precise applications of pesticides by smart machines or drones, automatic recognition of livestock health, and many other applications.

After analyzing the data and determining its insights, we use our clean data to train some advanced models to predict something or classify something. Say that we use the pictures of fruits to train the model which fruit is good and which is infected by some disease. Then we test the model on a fresh set of pictures that it did not see before to evaluate its performance. If it gets a satisfactory score on the test sample, we move to the next step, the deployment; otherwise, we repeat the training on more data or change the model hyperparameters to increase its performance. The deployment stage is done by coding the trained model to the corresponding machine, for example, a drone or a robot. Then we test the machine performance after deployment. Finally, we have a machine that can grade and sort fruits or detect infected fruits to discard them. A similar operation is repeated in every field of machine learning in agriculture.

Types of machine learning algorithms

There are different types of machine learning algorithms. These types are grouped into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

Reinforcement learning is used mainly in the robotics field, which is to let the model learn from its errors. For example, if we train an ML model to pick up ripen fruits, we release the picking arm to a specific distance. It tries to pick up the fruit. If it succeeds, the model can associate the apparent distance through its camera's input to the distance of expansion of the arm. If the robotic arm fails, it tries longer or shorter distances until it reaches 100% success. Learning from trial and error in ML is called reinforcement learning.

The supervised learning

In these types of problems, we have labeled data, i.e., the dependent variable is well defined, and we have one or more independent variables that are mapped to the dependent variable. For example, we have a dataset that determines crop properties and crop selling price. The independent variables are the physical properties like crop moisture content, sugar content, pH number, time since harvest, and others. On the other hand, the dependent variable is the selling price. The records of the dataset are collected from several farms for several years. This example is a type of supervised regression problem, as we can use regression algorithms to relate the price to the properties, then to use the resulting model to predict the selling price if the conditions are different. Several regression models were used to predict crop yield (109,110), crop detects disease (111,112), estimation of cattle weight (113), estimation of evapotranspiration (114). Other examples and applications are listed in (115), and more examples are listed in section 0.

Another example, if we have animals' vitals like blood pressure, body temperature, and the number of heart beats, and we have a column in the dataset named "Health Status", which contains two values, "Healthy" and "Unhealthy". This problem is a type of supervised classification problem. So, we can use some ML algorithms like support vector machine to map each set of vital signs to the corresponding health status. Classification models are used in many applications like classification of parasites of strawberry (111), weed detection (116), crop species recognition (117), Classification of cattle behavior (118).

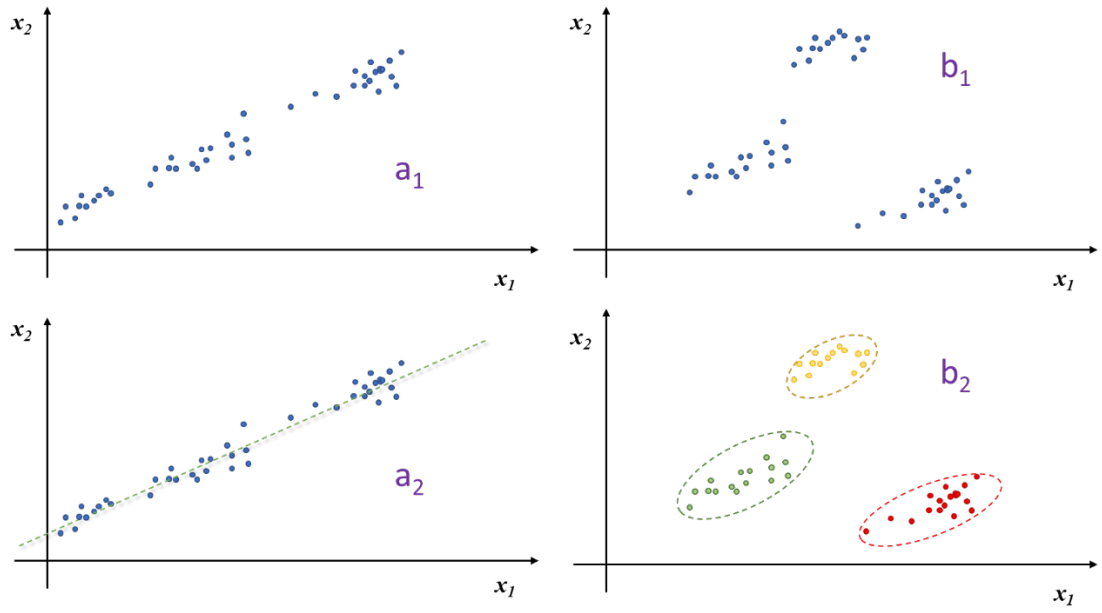

Figure 15 labeled vs. unlabeled data.

The unsupervised learning

Unlike supervised learning data, the unsupervised learning data are unlabeled. i.e., we do not know which datapoint belongs to which category unless we look at the data pattern. In Figure 15a, we can see a noticeable linear pattern of the data. Thus we used the regression methods to draw a regression line that we can use to predict the value of x_2 by knowing x_1. On the other hand, if we find data like Figure 15b_1, it is obvious that no trend controls the data, while there are some partitions of the data. The partitions or the clusters have standard features that are not necessarily obvious. Clustering should not be confused with classification, as the latter is a supervised algorithm, while clustering is an unsupervised algorithm. For example, we can have the vital signs and health status of some livestock. We can build a supervised learning model to classify the animal based on the given feature. This procedure is a tedious operation to have each animal examined to state his health status.

On the other hand, through the clustering model, we can plot the vital signs (even virtually if we have more than three dimensions). The clustering algorithm can specify which animal is close to which cluster, forming groups of points that have things in common. After the clustering, we can see the health status of each cluster. We can notice that one cluster contains the healthy animals, the other cluster contains the unhealthy ones, and the third cluster contains risky state animals. Later, when we have a new, unexamined animal, we can see to which cluster it belongs. The main benefit of clustering is that we specify the number of clusters before the modeling. We can try several numbers of clusters and measure the deviation between the members of each cluster. We can say that the best number of clusters to select is the number that leads to minimum deviation.

To understand the difference between clustering and classification, we can consider Figure 16. The figure shows a scattered dataset of some farm animals. If we applied clustering based on size factors and chose two as the target number of clusters, we got two categories, one shows livestock, and one shows poultry (Figure 16b). Alternatively, if we specified four target clusters, we could get what is shown in Figure 16c, which differentiate between small and big size livestock and between small and big

poultry. On the other hand, classification requires labels, Figure 16d, each type of animal is labeled. We train the model on a portion of the dataset, then test it on a smaller portion to ensure its accuracy. One of the benefits of clustering is that its results vary if we change the clustering features. For example, if we classify the same dataset on some physiological features, the rabbit will be clustered with livestock animals. They are all of a similar family, while the other cluster will contain only birds. Third clustering if we made the clustering based on marketing types in the USA, we would find camel, sheep, ducks, and rabbits in a cluster, while cows and chicken in the other. In contrast, the first cluster contains less consumable meat in the USA.

One of the types of unsupervised learning methods is the PCA or the principal component analysis. This approach is used for dimensionality reduction of models. Sometimes, the number of independent variables (features) is enormous, so the computer takes too long to train the model. The PCA can group several variables into one feature that carries most of the characteristics of the original variables. This approach can significantly reduce the number of features which helps reduce the amount of computation power needed for training the models and can save this time to enhance the model's accuracy. The PCA is used chiefly with categorical features like the text characteristics of the disease, where we can convert the text to numbers using one-hot-encoding. We apply the PCA to reduce the number of features (119). The one-hot-encoding method is used as follows:

Suppose we have some symptoms of some diseases for cows, such as shivering, fever, rash, irritation, loss of appetite, boils, and warts. Thus, we want to process a dataset with the symptoms as input, which is not possible as the symptoms are of categorical text format. We use the One-Hot-Encoding to convert the categorical column to numeric values that the ML models can analyze. If we have the symptoms column shown in Figure 17, the categorical features can be converted to numeric by assigning a column to every single feature, then giving the values 1 or 0 depending on the feature's existence in the symptoms column. As we see, the method converted the column to other readable columns by ML models. However, it converted one column to seven columns in this example, while sometimes the encoding could lead to thousands of extra columns. That is why we need the PCA to reduce the number of columns to a reasonable number.

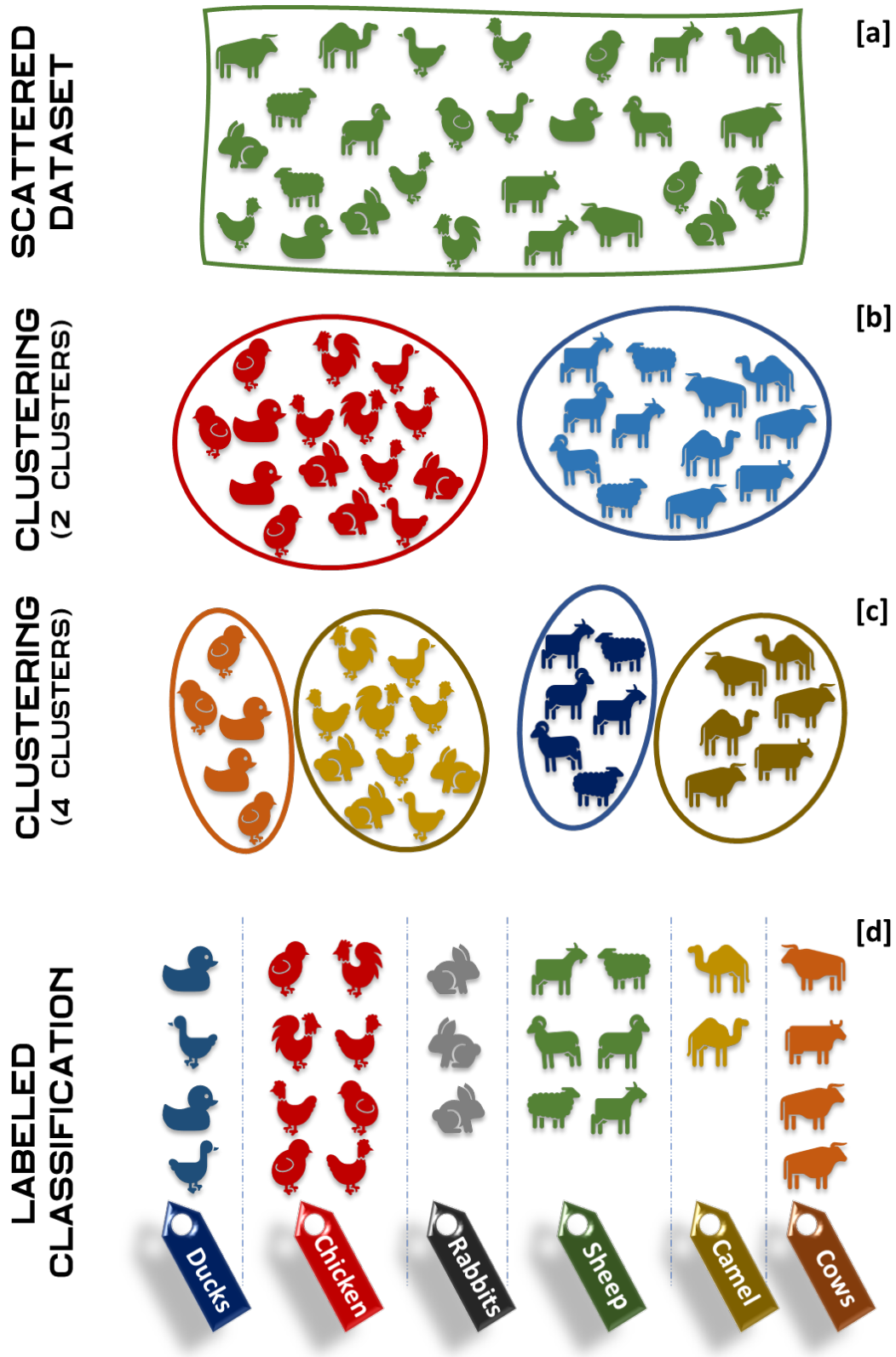

Figure 16. Difference between clustering and classification.

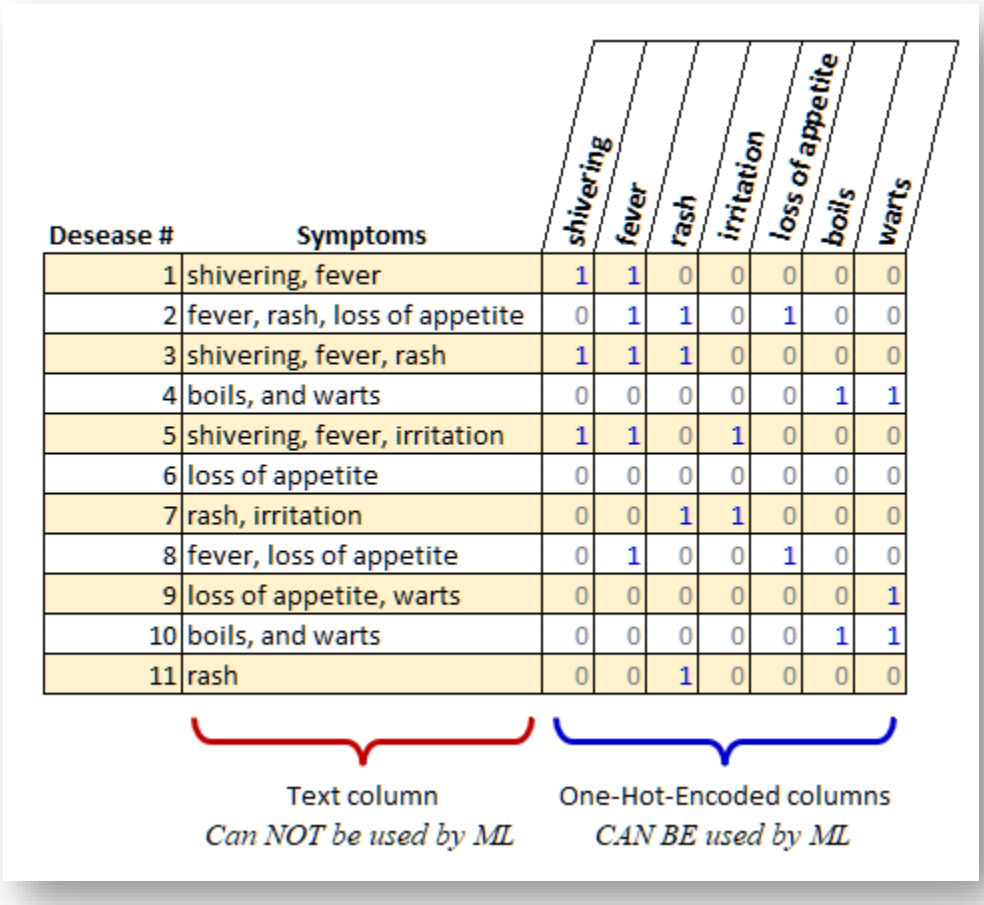

Figure 17. One-Hot-Encoding to convert categorical to numeric features.

Data Applications in the agriculture sector

Smart agriculture depends on sensors and data. We have mentioned many data applications that are used in the agriculture field. Here we will discuss the applications deeply to be better understood. The applications below include both big data ad conventional data applications.

Applications in soil and water

Soil science is one of the earliest fields that benefited from data science, especially in soil classification and hydraulic parameters estimation using neural networks like the work of Schapp and Leij (120). This work was followed by several works using machine learning techniques to estimate soil moisture (121–124), soil classes and classification (125–127), soil erosion (128–130), soil salinity (131–133), and soil mapping (134,135).

Machine learning also helped in solving many of the irrigation water problems, starting from predicting reference evapotranspiration (136), through assessing the irrigation water quality (137–139), modeling water infiltration in soil (140), managing and automating smart irrigation (141–143), and irrigation scheduling (144,145).

Applications in weather and climate change

Agriculture activities are highly affected by the weather. Determining the sowing date depends on the temperature expectations throughout the growing season, and the evapotranspiration calculation and the irrigation depend on several climatic factors. Thus, in addition to the mentioned works that used data science in predicting evapotranspiration, some works forecast weather changes to protect the plant from droughts (146–148), to find the optimal sowing dates (149,150), and for weather-based crop selection (151–153).

Applications in plant health and protection

Plant protection includes protection from viral, bacterial, and physiological diseases, in addition to protection from invasive species and weeds. Big data and machine learning applications helped in all fields of crop protection (154). Most of the detection methods depend on image processing models to detect the apparent symptoms of the diseases. These models can detect the disease and measure its severity (155,156). Similarly, weeds can be detected using image analysis of RGB and hyperspectral imaging with some machine learning classification models like random forest (157–159).

Applications in livestock and poultry production

Data applications benefited livestock farming and poultry production. Machine learning algorithms were used to detect livestock behavior like grazing, standing, lying, walking, and many other behaviors. The models can also detect strange behaviors that indicate signs of illness (160,161). Diseases on livestock are also being detected precisely and rapidly using machine learning applications (162–164). Other applications include classification of herd types (165), predicting body weight (166), managing herds using virtual fencing (167), improving livestock census data (168), and determining the nutrition requirements of livestock (169).

Poultry production also benefited a lot from data science and machine learning technologies. There are a lot of applications such as poultry welfare monitoring using cameras (image-based detection) (170,171) or by audio signals for early disease detection (172,173), poultry monitoring and behavior detection (174,175), eggs classification, and freshness detection (176,177), robotic picking up for floor eggs (178), poultry meat characteristics (179), prediction of meat yield (180), prediction of energy content in corn fodder (181), evaluation of the environmental adaptability of poultry housing (182), and detecting poultry eating behavior (183).

Challenges and obstacles

Using data applications for smart agriculture is promising. However, there are some obstacles in applying these technologies in a low-income region.

The first obstacle is the novelty of this field and the limitation of data infrastructures such as the farms equipped with sensors and modern techniques for data collection and transmission. This problem can be overcome by applying the results of ready-made models as a beginning and using the deployed commercial applications, such as the plant and animal health assessment software, drone monitoring software, and so on.

The second obstacle is the weak connection speeds or lack of internet in rural areas. This obstacle can be overcome by working on technologies that can run offline or rely on local storage. Different connectivity solutions can also be adopted, like ZigBee.

The third obstacle is the weakness of the human resources in dealing with modern data technologies, which can be overcome by conducting intensive training programs that attract distinguished youth and qualify them to work in the operation and development of data-based software.

The fourth obstacle is the legislation in data, where it must be clarified who owns the data and who has the right to use it. Additionally, it is essential to legalize the means of securing data and setting up legislation that sets out the penalties for those who violate this data, whether by illegally obtaining, it destroying it, or hacking its servers in any way.

The fifth obstacle is the financial and investment obstacles in data, as the field needs an expensive infrastructure. For this reason, investors must be encouraged to enter this field by informing them of its importance and highlighting the gains that they will get from applying data models in terms of cost savings and increased profits.