BOK CHOY PREDICTION MODEL ANALYSIS BASED ON SMART FARM USING MACHINE LEARNING

Aldi Sulthony Susilo[1], Dr. Nyoman Bogi Aditya Karna, S.T., MSEE.[2], Ratna Mayasari, S.T., M.T.[3]

[1,2,3]Telecommunication Engineering , School of Electrical Engineering, Telkom University
[1]aldisulthony@student.telkomuniversity.ac.id, [2]aditya@telkomuniveristy.ac.id,
[3]ratnamayasari@telkomuniversity.ac.id

Abstract

Indonesia is an agricultural country that has a dependency on the horticulture sub-sector. Bok choy is included in the mustard greens group as one of the strategic products from the horticulture. The needs for mustard greens are getting higher. Based on Indonesia's Central Statistics Agency data in 2019, the mustard beans production rate increases only 2.63 % higher than in 2018. If it does not meet the desired supply, it opens the possibility of a lack of bok choy supply at the market, resulting in high potential price fluctuations. These conditions initiate relevant system research to help the farmer develop a bok choy crop reference guide, especially in the seeding phase. In reducing the limitations caused by the lack of science and knowledge in the farmer environment, the prediction model is the proposed outcome by considering the use of IoT mechanism that has widely developed. The model is based on a system that integrates IoT's interest in the agriculture field, namely smart farm, for retrieving real-time data based on automatic control, MySQL database for storing data, and machine learning technique to establish the prediction model as the guide for the farmers to find appropriate parameters for planting bok choy. The prediction model performs using Python, a high-level popular programming language due to its ease and open source. Python interprets the bok choy growth dataset based on the irrigation system scenario from the integrated system with the relevant library of data preprocessing interest and the Decision Tree algorithm of the Scikit-learn library to train the model. The system conducts a series of machine learning phases to take the insight analysis needed to create a prediction model. This thesis's expected result is an ideal prediction model that results from the global system dataset based on the previous undergraduate thesis created as a reference guide for the farmer environment in planting bok choy during the seeding phase. The model performance metrics as the consideration in deciding the outcome model, which are accuracy and precision.

Keywords : IoT, smart farm, machine learning, Python, Decision Tree, Scikit-learn, dataset.

1. Backgorund

Indonesia is a developing country with a broad population segment that is depending on their life in agriculture. The horticulture sector produces vegetable products, fruit plants, ornamental plants, and medicinal plants that can be traded. Based on data obtained from the Central Statistics Agency in 2018, the number of farmers is around 33.1 million, and there are 10.1 million horticultural farmers [1]. As a result of this population, agriculture is still one of the Indonesian economy pillars by contributing 14% of the GDP [2]. Besides helping the national income, the horticultural sector also contributing to its export activities and food needs. Again, the farmers receive the annual revenues sufficient for their lives. The other parties is involving as the supporting system in food supply distribution apply before consumption by the communities.

One of the leading agricultural commodities in Indonesia is bok choy. This plant is required as a supplementary food material in the household and industrial environment. Bok choy is considering as one of the plants classify in the mustard greens group. Along with high demand resulting in potential supply and price fluctuations, the mustard greens production rate is only increased by 2.63 % in 2019 over 2018 in the Central Statistics Agency of Indonesia data [3]. The public information also reinforces the Ministry of Agriculture data states that there has been a fluctuation rate in the population of West Java province food consumption by day and year from 2013 - 2018 [4]. The fluctuation is caused by several factors, including weather sensitivity, limited resources (producers, planting knowledge, infrastructure, and land availability), the influence of fertilizer, and climate change. In maintaining the stability of bok choy commodity prices, an adequate supply of the finest quality of bok choy is needed. The specific observed parameter to analyze is the soil moisture as the growth-related factor to plant the bok choy.

The objective is increasing the production of bok choy, which offers to farmers environment as the main player in producing the best choice of bok choy seeding phase in increasing the possibility of optimal crops and open opportunities to maintain the availability of bok choy quality on the market. One of the facilities used to do suitable planting is the greenhouse. The greenhouse is an architecture to plant the bok choy following the

desired optimal condition with a more attentive environment and minimize unwanted environmental factors. In addition to utilizing the environment independently, the researchers have conducted various researches on the use of greenhouses, including real-time monitoring and controlling greenhouse systems based on smart farms system [5]. An IoT device is connected through a network as an integrated system in retrieving real-time data is necessary. Raspberry Pi is used as an IoT device because it has an embedded wi-fi module and external instrument, such as sensors as a supporting system. The soil moisture sensor is defined as calibrating the water dose value of greater than or equal to 15 % as the lower limit and lower than equal to 25 % defined as normal condition [6]. Whereas the optimum temperature is in a range of 20 - 25 degrees Celcius [7]. The application of Raspberry Pi in the greenhouse as smart farm system implementation has been widely used by researcher [8][9].

Users can perform data retrieval results for a prediction model to determine the best composition in producing optimal bok choy seeding crops. This method has been proven by several researchers who made a collection of observation data results in tomatoes' and lettuce growth as it used for quality predictive models [10] [11]. The Raspberry Pi is retrieved data from the respective types of desired sensor monitoring and controlling and is automatically stored in the database. The platform is supported with the additional application of the database management system field, such as MySQL database. The designated raw data is divided into several groups. The group of data combines the dataset as the input for performing the prediction model. This integrated system of IoT and machine learning has also been explored by the scientist [12] [13].

A prediction model is generated using the machine learning technique in processing information from the available data sets split by the irrigation system scenario. ML performs with the help of complementary libraries, such as Numpy, Pandas, Matplotlib, Seaborn, and Scikit-learn of Python programming language that has been widely used in the related field to build the learning systems in producing a desired model [14]. A classification approach of supervised learning applies to develop the model strengthen by the designated output label that manually inputs. The model development is conduct in an open-source Jupyter notebook as it is interactive to the user to conclude the analysis of each machine learning phase. There are several procedures as a reference to perform the model testing. The train/test split procedure splits the dataset automatically and implements a decision tree algorithm to train the data. The evaluation of the model performance performs with three classification metrics, such as accuracy score, confusion matrix, and classification table that correlated to each other as the analysis to determine the prediction model's quality. The model is saved to the local computer model for repeatedly implementing the model testing, which applies to the pickle library's given input.

In this thesis, a designated system aims to generate an ideal prediction model of bok choy growth crop, especially in the seeding phase. This research uses a machine learning technique to develop the prediction model, adopts an integrated system that has been established for creating the dataset. The smart farm system and MySQL database for data storing is discussed in the previous thesis published internally in the university field [15] [16]. This thesis design discussion is divided into five chapters, from the introduction to the conclusion and suggestion section.

2. Basic Concepts

This chapter contains the definitions and basic concepts of the methods is used to design this thesis.

2.1 IoT

IoT is related to a small physical device equipped with a built-in wifi module and sensor, an integrated computing system. One of the IoT device is Raspberry Pi [8][9]. The picture of Raspberry Pi is seen in figure 2.1. Raspberry Pi 3b is amongst the popular has fundamental use in providing storage for data sets. MySQL database is capable of being installed as an external data management system. Python has also been the preferred programming language to execute coding in the device operating system, called Raspbian. Because of its many advantages, Raspberry Pi is classified as a minicomputer.

IoT retrieves real-time data with the sensor's help due to related factors on the object being monitored during its operation. The IoT's ability to send the gathered data over a network without any human to human or computer interaction is one of IoT's objective to accommodate the automatic operation [17]. This ability achieves with the help of a built-in wifi module that embeds in the device. IoT device develops because of their role, which helps humans control without boundaries in time and distance. Besides, the IoT benefit could save costs and reduce human resource problems in retrieving more accurate data [11]. IoT personifies the ease of human life is implemented to provide solutions to the issues in agriculture [12].

2.2 Smart Farm

It is an specific area of IoT implementation that is widely used [5] [12]. IoT platforms cover the intelligent system to implements in the field of agriculture. The term" smart" is the origin well-developed of the Internet of Things in leverage of real-time event. This system is expected to help the farmer to be able to create a solution based on the farming field problem of growing plants, which are influenced by

Figure 2.1 The IoT device

uncertain environmental and growth-related factors. The system deployed in the farm could control and monitor the plantation that impacts the observed plant to identify better and understand the pattern related to the treatment given. The specific environment is correlated also retrieves the data to store as the datasets. The datasets as the input data to accomplish a comparative analysis of the hard-predictive information of the plantation [5].

2.3 Machine Learning

Machine learning is a subset of the artificial intelligence field besides deep learning. The focus of ML is concerning the machine's capability to study the insights possible from the assigned data or environment in developing the desired prediction model [9]. Predicted data consider the output of this area. The machine learning research area is implemented in most real-world cases; one example is agriculture advisory [18]. There are three types of most discussed problems designated in this interest: regression, classification, and clustering. It differentiates by the given structure distribution of the variable corresponds to its attributes and the defined output availability in the dataset [19]. This research area divides into four techniques based on the characteristics of dataset contents, as in figure 2.2 and discussed below:

2.3.1 Supervised Learning

The technique that applies to the dataset, which has labeled data. Regression and classification are the types of problems classified using this technique differentiated by the observation data types. Regression handles continuous attributes, and classification takes categorical attributes. In this type of machine learning technique, the dataset includes with desired target output. Then, the suitable algorithms seek to learn the patterns involved in the dataset by the specified input. This learning generates the predicted data corresponding to the associated parameter of the specific class [20].

2.3.2 Unsupervised Learning

A type of learning technique applies to the dataset provided with unlabeled data. Clustering is one of the main problems categorized in this learning. The detailed input data give guidance to suitable algorithms to detect and gain rules of the related patterns. The essential insights are then summarised to provides knowledge improvement to generate the predicted data. The dataset of this technique application does not include the target output variable [20].

2.3.3 Semi-supervised Learning

This technique applies to the dataset, which covered a combination of supervised and unsupervised learning data [9]. The supervised learning comes with labeled data. On the contrary, unsupervised learning comes with unlabeled data. Subsequently, the dataset applies for this learning arises with a little of observation's labels, while the rest is filled without labels. This procedure exploits the dataset's structure patterns even though the observation's target output is not available. The technique sustained with a little information about the target output label associated with labelled data to the related parameter of unlabelled data [21].

2.3.4 Reinforcement Learning

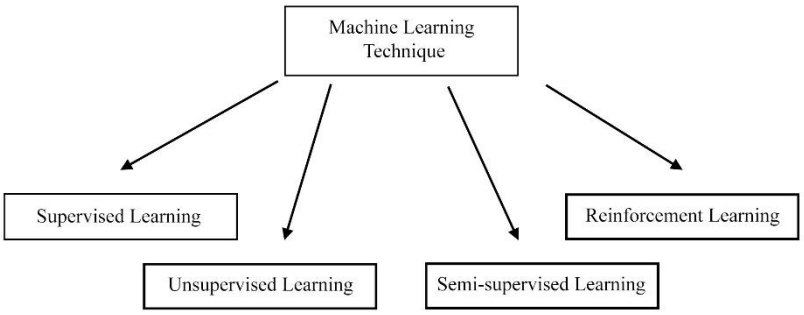

Figure 2.2 Types of machine learning

An approach that allows the machine to perform technique by observing the surrounding environment leads to feedback, either rewards or penalties [22]. It is familiar as a reward-based learning system [9]. The aims of finding patterns in the external environment to determine the ideal action depend on the availability of a specific observation parameter for performance improvement. Reward or penalties feedback the machine uses that to learn the patterns.

2.4 Dataset

Dataset is a final compilation of data retrieved from an observed system through an information source, such as a device or sensor [11]. The dataset is visualized with a tabular form in various data formats, such as NumPy, Microsoft Excel, CSV, SQL database, Etc. There are two types of datasets used for machine learning application, train and test data. The train data has a function as primary data for creating a model. Applying a suitable algorithm to learn concept applies to the related algorithm and gives the model outcome. In contrast, the test data can evaluate the algorithm condition after the training data applies. The dataset contents determined by the attribute and the designated label correspond to the appropriate approach [22]. Every dataset has its characteristics. These are the differences discussed below:

1. Variety

   This phrase refers to the diversified data applied to the system [22].

2. Veracity

   This phrase represents the data aspect, such as noise, abnormality, etc. [22].

3. Volume

   This phrase points out the number of data composed at a given time [22].

The information represents corresponding to its axis, horizontal, and vertical in the dataset table's contents. The information present in each horizontal axis divides the column header, and each row corresponds to its attributes, class, and instances. On the contrary, the information present in each row is called the variable or value [19]. The explanation of each dataset contents term is shown in figure 2.3 discussed below:

1. Variables

   It corresponds to the data aspect captured for the observations. For example, the rows represent the number of plants, and the columns contain the characteristics of the plant, such as soil moisture, plant height, Etc. Its also known as the value type of the attribute. There are various types of variables available for the observation, such as nominal, ordinal, integer, interval-scaled variables, Etc [19].

2. Attributes

   The variable associated with the same relationship corresponds to its resources. There are two types of attributes that are widely used. There are categorical and continuous. Nominal and ordinal corresponds to categorical attributes value. In comparison, integer and interval-scaled variables corresponds to continuous attributes value [19].

3. Class

   It is the exclusive connotation for one attribute contains the designated target outcome. This case

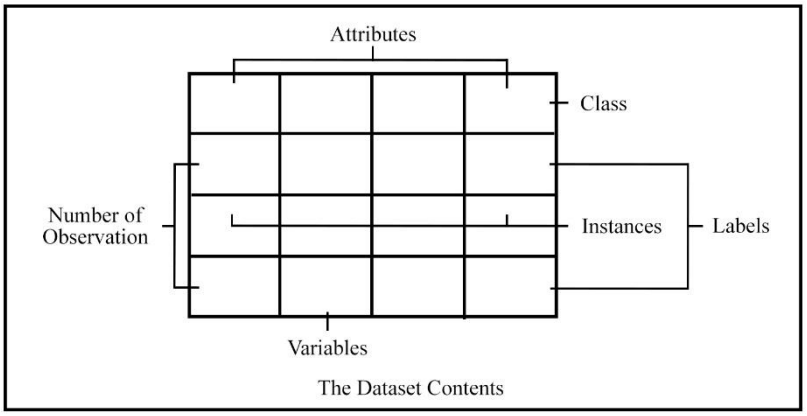

Figure 2.3 The dataset contents

available only for the labeled data, in which the class column has an assigned value [19].

4. Instances

   Also known as the dataset example corresponds to a collection of variables belongs to each observation [19].

2.5 Jupyter Notebook

It's a type of open-source project notebook. It was established in 2014. The notebook is life-time free of use under the restriction of the BSD license. The objective is contributing scientific computation in the data science field that supports all programming language environment, mostly Python. The notebook provides a wide variety of related library used to perform any desired model. This notebook helps by the immense number of contributors and community in GitHub [24].

2.6 Python

Python is an open-source programming language. It is suitable to operate based on the widely used operating system, such as Windows, Linux/UNIX, Mac OS X, Raspbian, Etc. This language is continuously developing, which there are two current versions available, version two and three. The differences are located in its syntax operation. The use of Python is increasingly massive, especially in the field of machine learning research area by provides many free functions corresponds to the related libraries and packages [25]. This language is an interpreter where the Python reads one line of code after being interpreted into a common machine language and then will be executed. The discussion of related libraries for specific purposes in the machine learning area is explained below:

2.6.1 NumPy

Numerical Python is a library, besides SciPy, for scientific and vectorized computing to perform other advanced libraries. This library has a large number of contributors. The library is suitable for performing various mathematical disciplines, such as generating N-dimensional array object, linear algebra, Fourier transforms, and random integer number [13] [25]. The beneficial function is capable of integrating with a wide variety of databases, and the code is stable [25].

2.6.2 Pandas

A library provides dataset preparation concern that does not have any strict competitor among its interest. Several types of problem, such as cleaning and manipulating of the data [13]. This NumFocus sponsored project provides many functionalities choices for integrated dataset synchronization and handling from various input/output data formats, such as Python/Numpy, Microsoft Excel, CSV, MySQL database, and HDF5 data format. This library has implemented operation, such as easy to influence disorganized data to relevant data by unique indexing and subsetting with label-based and integer-based slicing to attach specific dataset information [25].

2.6.3 Matplotlib

One of the Python libraries for performing data visualization capabilities besides Seaborn and Plotly [13]. This library is written on a low level with many possibilities for plot customization. Visualization of the dataset by this MATLAB-based plots with numerous plot schemes is possible, such as line, scatter, histogram, and others, depending on project interests. To master its central concept in visualizing various plot design, learning the fundamental idea is the best recommendation for deep

understanding. There is a lot of syntax operation to conduct a series of the custom graph.

2.6.4 Seaborn

A high-level depiction library for user-interactive statistical graphics, as one of the concerns in data visualization phase. Seaborn is built on top of Matplotlib but still has limitations over Plotly in powerful plots documentation [13]. This library is suitable in manipulating patterns and grids makes it more comfortable creating existing sophisticated detail plot. Still, it has the constraint in the variety of plot customization that Matplotlib does. Seaborn makes users intelligent in data visualization with word efficiency and various colour palettes for more thorough insight [25].

2.6.5 Scikit-learn

A Python library that focuses on machine learning applications. This library offers various implemented machine learning categories and its supported algorithm that other competitor libraries, such as Shogun, Mlxtend, and Mlpy, do not fully provide based on the number of implemented algorithms in related categories [13]. For example, Scikit-learn provides the optimized performance of regression and classification problem approach, such as decision tree, support vector machine, and logistic regression [13] [17]. The popularity provides by this library is continuously present, wherein most applicable real-world cases, such as Github, a Microsoft subsidiary web-based hosting for software development and version control, by using the number of projects, releases, and contributors [26]. Also, Kaggle, the largest data science community website, uses the number of notebooks, datasets, contributors, and release date [26].

2.7 Decision Tree

One of the wide range of methods for performing the machine learning technique is developing a prediction model based on the given dataset. This method is suitable to conduct classification and regression problems in the form of a set of decision rules [19]. The decision tree performs by a splitting data procedure based on attribute value in the dataset until each corresponding branch is perhaps labelled by one regression or classification attribute. Corresponds to the issues, the decision tree method is possible to interpret two types of the tree. The classification tree correlates with the categorical decision variable. In contrast, the regression tree corresponds to the continuous decision variable. This algorithm offers various advantages, such as competent in data missing values handling as one of the data cleaning procedure with the most suitable value an offers an efficient large performance of tree traversal algorithm. Also, generating straightforward decision rules that ease of interpretation [12] [18]. This algorithm proposes to give the prediction outcome in the type of classification of other unseen instances. The given new input, the collection of the unseen instances is known as a test set, or unseen test set different from the original training set. In predicting the class label of unseen instances, The mechanism of the decision tree is started from the root node [12]. The label's attribute validates each attribute's value split until the leaf node corresponds to the relevant parameter. The terminology is discussed below:

1. Root node

   It is a phrase that corresponds to the original training set as the parent node. This node is splitting into two or more possible sets.

2. Child node

   This phrase refers to the node based on the root node splitting result.

3. Internal node

   This remark refers to the tree's content not classified as the root or a leaf node.

4. Decision node

   It corresponds to the sub-nodes split into more sub-nodes as a result.

5. Leaf node

   This phrase points out the bottom nodes that are the final level of the node. This node corresponds to the instances of a training set subset that resulting in the same classification outcome.

6. Branch

   This phrase refers to the entire decision tree section.

7. Splitting

It is a procedure that refers to split a node into sub-nodes.

8. Pruning

   It is a procedure that refers to the reverse process of splitting, in which discarding the unrelated sub-nodes.

2.7.1 TDIDT Algorithm

It is known as the principal approach of decision tree implementation also as the fundamental knowledge of other notable classification algorithms, such as ID3, C4.5, and CART [19] [28]. This algorithm requirement's is specifying the categorical attribute's value of the dataset example. The specification is based on the elected of attribute splitting in each of the non-leaf nodes. The essential qualification of the TDIDT algorithm is related to control flow mechanism, in which it operates the pattern below." IF the dataset example corresponds to the same class THEN the classes value is return" The TDIDT algorithm has a prerequisite before its implementation. It has to meet the sufficiency condition. A case that assures the training data is consistent, 15 where there are no two instances of all attributes value is associated with distinct classes. The algorithm's operation depends on composing the best choice of attributes to split in every stage. In this type of algorithm, a primary issue is still present. The problem is there is not any guideline available that point out the split selection of an attribute [19].

2.7.2 CART

It is known as the abbreviation of classification and regression tree. This algorithm supports the decision tree classifier in the Scikit-learn library. The basis of this algorithm is the TDIDT algorithm. The algorithm supports continuous variable as the target variables, and it's not performing finalize complete rules [29]. This algorithm is well-known because it is operating in a wide variety of application. Also, this algorithm as a shortened way to avoid the over-fitting of data. There are two-stage performed, such as the evocation and pruning of the decision tree. Define the Gini index minimum as the parameter in the decision tree classifier function to select the sufficient attributes. This operation performs to completing the binary tree [30].

2.8 The Performance Metrics

There are three types of commonly used metrics for evaluating the classification approach. The analysis of the metrics is discussed as follows:

2.8.1 Accuracy Score

The type of metrics to update the model evaluation is based on the accuracy value. The essential outcome is resulting in a float number. This evaluation parameter derives from the real positive and true negative instances as it was also available in the classification report content. The function of these metrics is to make intuitive delivery of information [19].

2.8.2 Confusion Matrix

The types of summary to publish the performance measurement of a classifier in the prediction model. Classifier performances divide into two types of instances, positive and negative. Each category is depicted in a tabular form corresponding to the direction. The classes in the table's main content are actual and predicted. Four fundamental parameters support each class. The function is to demonstrate the instances frequencies based on correctly classified or misclassified data [19]. The discussion about each fundamental is explained below:

1. True Positives

   It is located at the intersection between actual positive and positive predicted class, explaining the total positive instances that are correctly classified as positive [19].

2. False Positives

   It is located at the intersection between actual negative and positive predicted class, explaining the total negative instances that are correctly classified as positive [19].

3. False Negatives

   It is located at the intersection between actual positive and negative predicted class, explaining the total positive instances that are correctly classified as negative [19].

4. True negatives

It is located at the intersection between actual negative and negative predicted class, explaining the total negative instances that are correctly classified as negative [19].

5. Positive

   It is located at the intersection between the total instances column and the actual positive class, explaining the total number of positive instances by adding true positive and false negative examples [18].

6. Negative

   It is located at the intersection between the total instances column and actual negative class, explaining the total number of negative instances by adding the false positive and true negative instances [19].

2.8.3 Classification Report

A performance measurements summary of a classifier based on the respective test set in which positive and negative instances are fixed. There is no limitation of any classifier used. Various performance measures are used to analyze the seven observed classifiers that characterize by the existence of true positive and false positive rate, which derive all other measures [19].

2.9 The Perfect Classifier

In this case, the instances' total is the absence of misclassified, which True Positive is correctly classified as Positive and True Negative as Negative, respectively. To quickly understand these cases' concept, the confusion matrix is derived in table 2.1 drawn in a tabular form. The performance measurement calculation operates through breakthrough analysis by using the formula given in the previous chapter. The value of True Positive Rate is one, in which the TP Rate is expressed by P divides by positive instances or P. Its value is the same for Precision, f1-score, and Accuracy measures. The differences in the output locate at False Positive Rate, which yields a value of 0. FP Rate influences 0 divides express by the subtraction of 0 with the negative instance or N [19].

3. Discussion

This chapter contains the proposed bok choy growth prediction model and system.

3.1 The Workflow of the Global System

The proposed global workflow system is started from retrieving data stage with IoT mechanism, stores data in MySQL database, and develops the prediction model. The data is sensed by the IoT platform with the help of the four respective sensors, which categorized into the value of five attributes given are as follows:

1. Webcam retrieves the picture of bok choy seeding period of the entire pot taken once a day at noon.
2. DHT22 sensor retrieves the room humidity and temperature value.
3. BH-1750 sensor retrieves the room light intensity value.
4. YL-69 sensor retrieves the soil moisture value from each pot.

The sensor retrieves real-time information based on the sensor's ability taken in a greenhouse located in Buah Batu region, as seen in figure 3.1. The exception for the soil moisture outcome, because it has a

		Predicted Class		Total Instances
		Positive (+)	Negative (-)	
Actual Class	Positive (+)	P	0	Positive
	Negative (-)	0	N	Negative

Table 2.1 The confusion matrix of perfect classifier

Figure 3.1 The greenhouse

different scenario based on the irrigation system, categorized into automatic and manual. The automated system is a process to provide water to the bok choy plant that has been programmed with relay help, which will irrigate automatically if the soil moisture detects the water amount in the observed plant is less than is needed. In contrast, the manual system is a process to deliver water to the bok choy plant by hand-operated, which is self-irrigate the plant based on the range of the time decides, 09.00 A.M. to 11.00 A.M., is the best time to irrigate the plants. The soil moisture value has been converted from a percentage to the categorical attributes at receiving data. The greenhouse as the plant placement used in this thesis.

The sensed data is sent to the Raspberry Pi 3B. The device gathers and stores the information that it retrieves based on the sensor placement. The sensed data is then sent through the localhost network automatically every ten minutes to the MySQL database as data management for complete store data. In the database, the sensed data divides into six groups, which are:

1. Camera's group contains the picture's path of the entire pot.
2. Plant A's group contains date-time and soil moisture value from the observed plant.
3. Plant B's group contains date-time and soil moisture value from the observed plant.
4. Plant C's group contains date-time and soil moisture value from the observed plant.
5. Plant D's group contains date-time and soil moisture value from the observed plant.
6. Room Condition's group contains date-time and greenhouse humidity, temperature, and light intensity value.

These database contents are seen in figure 3.2. The database has been deployed to website hosting for monitoring purposes. The dataset's scratch obtains from the 24 MySQL database has been prepared by merge mechanism corresponds to each experimental plant group and the room condition group that performs in the database by using MySQL query. Figure 3.2 The database overview as the data storing management used in this thesis. There will be two datasets generated as the input data to develop the prediction model based on the Excel file export mechanisms. The machine learning technique is

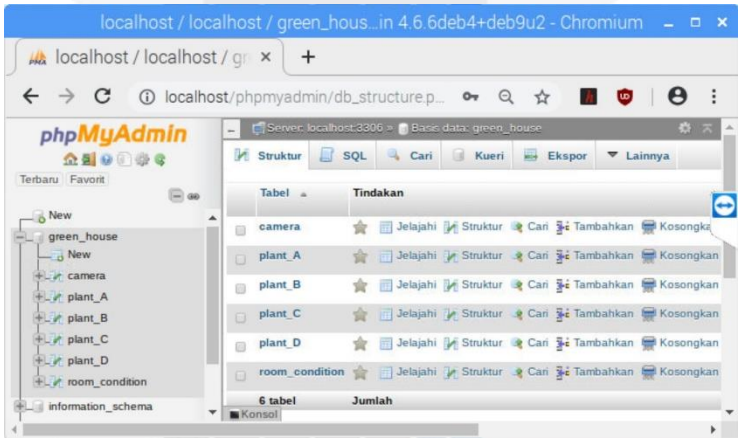

Figure 3.2 The database overview as data storing management

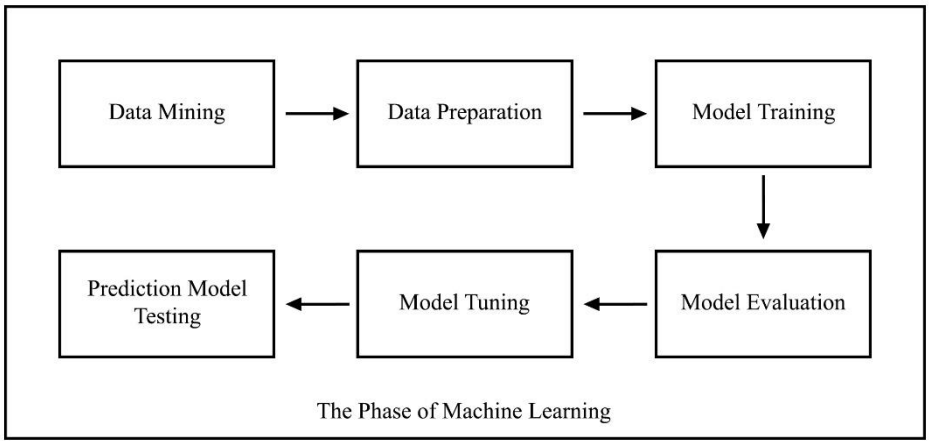

Figure 3.3 The proposed design of the prediction model

implemented after the dataset file preparation phase, with a suitable machine learning algorithm, achieving the prediction model.

3.2 The Workflow of the Prediction Model

This thesis's expected output is the prediction model using the system discussed in the previous section of this chapter. The prediction model design workflow is divided into five phases; these are machine learning phases visualized by a flowchart, as shown in figure 3.3. The operation consists of loading the data groups and adding relevant growth-related factor attributes to establish a dataset based on the irrigation scenario. The data mining phase outcome is the dataset that has labeled data in the class column. The subsequent operation is preparing the dataset to perform a machine-learning algorithm to obtain a model and analyze its performance metrics. Also, visualize the information given from the preparation and analysis stage as part of the notebook's storytelling in developing the model. This process is simply for intuitive reading, and easy to compare each variable to its attributes as gaining more reliable knowledge.

The dataset contains a continuous and categorical attribute also the desired class as the label in the class column for the target output variable. The machine learning technique used is supervised learning detailed by the classification approach based on the dataset's content. Machine learning conducts a series of procedures to gain insight from the dataset to achieve the outstanding model. Three characteristics

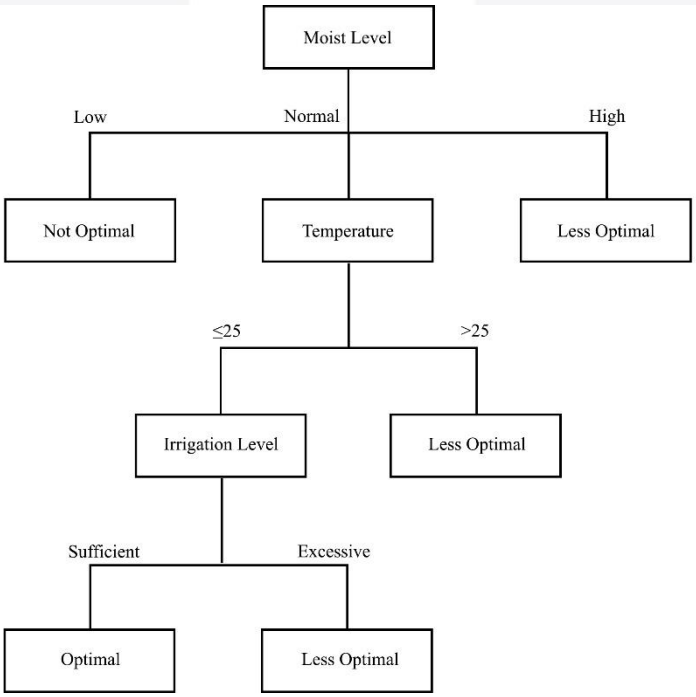

Figure 3.4. The overview of the classification tree design

influence it: variety, volume, and veracity of an attribute dataset. Besides, the number of data also plays a vital role in resulting in the ideal prediction model. In data mining, the focus is to retrieve the plant growth-related group corresponding greenhouse environmental data from the database. The data preparation phase performs the insight exploration of the data. This stage also conducts concatenating the final dataset from the two scenarios of irrigation system dataset with discarding the camera's output. NumPy for scientific computation and Pandas for data preprocessing stage is performing.

The Matplotlib and Seaborn are also conducting to make the easier interpretation as the comparative analysis of the attributes knowledge to develop the prediction model. The third phase is the crucial part, which is the model training. The dataset separates by using the train/test split procedure. Then, the train data trains with the Decision Tree of the Scikit-learn library. This library base is the CART algorithm. The algorithm provides the classification tree scheme of the defined attributes to split by using several command implementation parameters. The classification tree is shown in figure 3.4. Besides the train data, the test data is also available to perform the classifier. This phase's final operation performs the classifier using the test data corresponding to the training process variable, namely the classification tree added with the Predict function. The overview of the overall model training phase is shown in figure 3.5.

The model is completed after it generates the classifier. Evaluate the prediction model by using performance metrics. The metrics used are the accuracy score, confusion matrix, and classification result, which correlated. Another two metric performs with the different visualization of present the result. It is visualized with a tabular form. Then, perform the classification result to gain additional important parameters that support the associated parameters. Observe the parameter of the accuracy and Precision of the prediction model. Determines the model has achieved excellent condition. If it needs to develop more to accomplish the best shape, the model will direct to the beginning of the data preparation or model training phase. This process is not achieved automatically but by the author manually by hand-operation the data preparation phase's introductory command.

The test phase is conducted after the model evaluation is satisfied. Its objective is to create a prediction based on the model saved in the preferred directory after performing the model training phase with Python's Pickle library. The tree models correspond to its usability, namely, the trained model with Decision Tree classifier, Moist Level encoder, and Irrigation Level encoder. The most critical phase that is done next is to define the favoured unseen instances to test the model. The prediction model function is the best choice so that input can be given directly according to the variable repeatedly.

3.3 Diagram Blocks of the Prediction Model

The diagram blocks are seen in figure 3.6. It is started by compiling the dataset in the data mining phase. The data preprocessing process is performed in the data preparation phase in Jupyter Notebook to conduct a machine learning model. The dataset is separated into the train and test set. Each set has a process flow, especially for the training data, in which using the decision tree algorithm trains the train set. In contrast, the test data does not have any process to apply. This process is done individually, starting from the train then test set. The group is combined to achieve the classifier is performing the model train phase. Then, the model evaluation serves to identify the performance of the model. The performance metrics

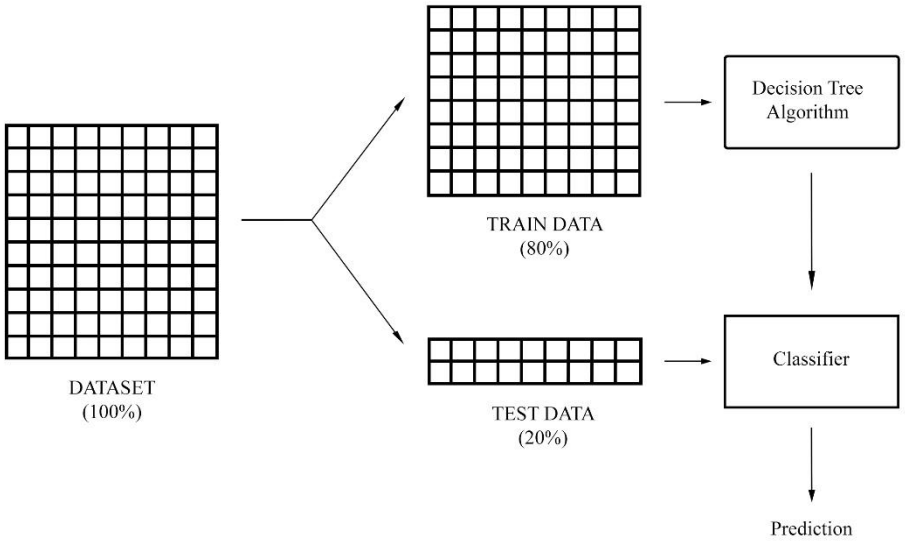

Figure 3.5 The proposed train and test procedure for model training

evaluate the model. There will be two conditions generated by the model, ideal or not ideal. If the result of the accuracy and precision value is satisfied, then the system is finished. Otherwise, it will enter the model tuning phase. The procedure started at the beginning of the data preparation based on suitable procedure needs. The system stops until the result of the prediction model is ideal. Perform saving the model to conduct model testing in another Jupyter Notebook with the preferred output corresponds to the available attribute. Then, the process of developing the prediction model is finished. This procedure operates hand-operated individually by executing each corresponding command.

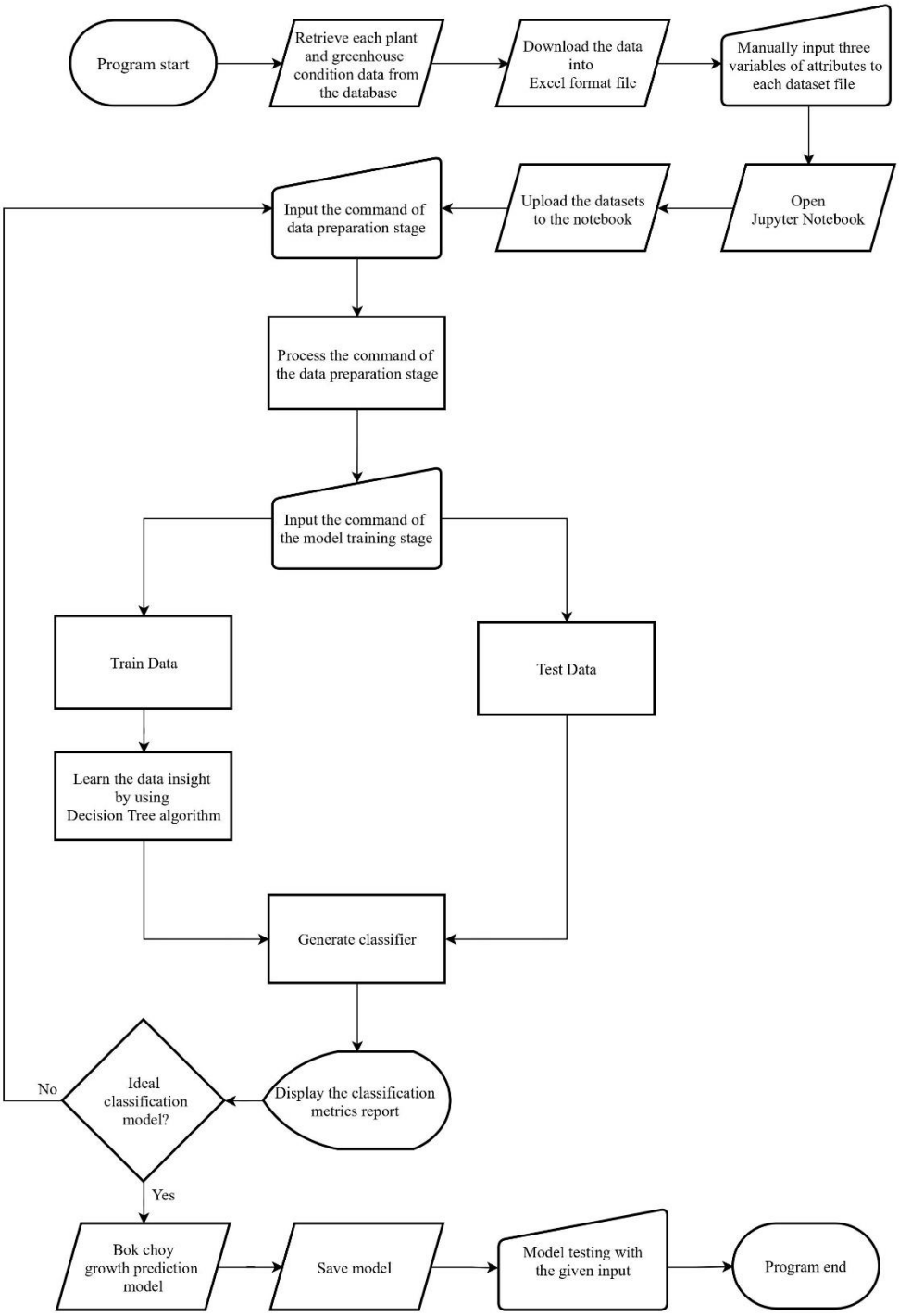

Figure 3.6 The proposed diagram blocks of the prediction model

4. Result and Analysis

This chapter contains the proposed bok choy growth prediction model outcome analysis.

4.1 Measurement of Download Dataset from Database

The measurement conducts in the Database Data Preprocessing Python Notebook to retrieve the available dataset The result of retrieving data process from the database is shown in figure 4.1. There is four connection established from the local computer to the database, in which done separately based on the order. Each of the connections represents the plant raw data group merged with the respective room condition. The resulting product is in the form of array data that corresponds to the corresponding sensor's value.

4.2 The Evaluation of Classification Performance Metrics

The performance metrics evaluation performs based on the available type for the classification approach. The corresponding value is the label test data, and the label predicts data based on the prediction made for the attributes of train data. The variety of classification metrics is discussed below:

4.2.1 Accuracy Score

The analysis of this type of metrics informs the accuracy score besides the classification report content that unites another parameter. The result of the accuracy score of this model is shown in figure 4.2. The actual outcome is in the form of the float data type. Nevertheless, the conversion to the percentage used in this thesis is to make intuitive information on the metrics. As one of the parameters includes the classification report, this value is based on the resulting total of TP and TN of the confusion matrix. The further discussion of this result review discusses in the classification table.

4.2.2 Confusion Matrix

This type of metric is necessary to calculate other performance parameters in the classification table, such as Accuracy, Precision, Etc. The result generated from the Jupyter Notebook does not give any remarkable information corresponds to the value. In creating a more straightforward interpretation to understand the outcome, the outcome value place in the table that has been discussed in chapters two and three. The visualization of the confusion matrix with a table is shown in table 4.1.

The noticeable perception is to give attention that the left top of the table corresponds to the True Positive value, and the left below is related to the True Negative class. The table means the classifier has correctly classified the total of positive and negative instances. This matrix informs the total instances for two-class in which is stated by the overview of the classifier's prediction discussed in the previous section. This type of confusion matrix is correlated with the Perfect Classifier Performance. It is strengthened by the matrix's pattern with no value correlated with the False Positive and False Negative. As confirmation, it is suitable to review the parameter correlated. The TP rate expressed by the total of True Positive instances divides by the total of positive instances equals one is to meet the criterion as a result is equal to one. False Positive rate equals to zero as there are no available false positive instances classified. Also, the Precision equals one as the positive value has only the True Positive value.

```
Connect to database green_house on 192.53.118.245
Connect to database green_house on 192.53.118.245
Connect to database green_house on 192.53.118.245
Connect to database green_house on 192.53.118.245
```

Figure 4.1 The data retrieving result from the database used in this thesis

4.2.3 Classification Report

These metrics perform as the final analysis as the value derived from the confusion matrix. The information is present in the no-line tabular form is seen in figure 4.3. The performance parameter as the consideration in depending on the ideal prediction model is the Accuracy and the Precision. Before reviewing the parameter based on the value derived from the confusion matrix, these metrics clarify that the noticed label is only two, namely Optimal and Less Optimal, as it is not clearly stated in the confusion matrix discussion section. The absence of the Not Optimal is because there is not enough instances available in the dataset. The accuracy outcome meets the perfect classifier case prerequisite requirement as it also clarifies the accuracy score that has been converted into the percentage. Either precision score or recall is also completing the qualification. The other available parameter in the classification report, namely F1-score, is not analyzed further because their basis to measure is based on Accuracy and Precision. Another parameter, namely Support, corresponds to the number of TP and TN, respectively.

5. Conclusion

A well-work of the global system is achieved considering creating and using dataset for developing the prediction model with a not that this system has not been integrated automatically. The dataset's content is using the classification approach of the supervised learning. There are two datasets available based on the irrigation system scenario. However, the prediction model performs is used the concatenating of the datasets. NumPy is worthy of performing the scientific and vectorized computing-based on the continuous attributes in the dataset. Pandas is satisfactory to perform preparation stage to explore the dataset insight. Matplotlib and Seaborn is appropriate to perform the visualization of dataset insight as comparative analysis of the dataset attributes value. The Decision Tree algorithm of the Scikit-library is suitable to develop the prediction model in this thesis with preprocessing stage of the categorical to the continuous attributes with label encoder. The prediction model performance is classified as the perfect classifier case at it has only the TP and TF values with no correlated value corresponds to the FP and FN. There are only two labels noticeable of the prediction model by performance metrics as it only has very little data that is also strengthened by the data visualization. The accuracy and precision parameter outcome is satisfied resulting in 100%. The model predicted only satisfied the

```
# Import a mandatory library to perform model metrics report
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Inform the result of the accuracy score result
acr_scr = accuracy_score(y_test, y_predict)
print(f"Accuracy Score: {acr_scr * 100:.2f}%")
```

```
Accuracy Score: 100.00%
```

Figure 4.2 The overview of the accuracy score

		Predicted Class		Total Instances
		Positive (+)	Negative (-)	
Actual Class	Positive (+)	313	0	Positive
	Negative (-)	0	6	Negative

Table 4.1 The visualization of confusion matrix table

```
Classification Report:
              precision    recall  f1-score   support

  Less Optimal       1.00      1.00      1.00       313
       Optimal       1.00      1.00      1.00         6

      accuracy                           1.00       319
     macro avg       1.00      1.00      1.00       319
  weighted avg       1.00      1.00      1.00       319
```

Figure 4.3 The overview of classification report result

Less Optimal label as it has a lot more data than the Optimal label data with the value of three hundreds and thirteen by six instances.

References :

[1] C. S. A. of Indonesia, Results of Inter-censal Agricultural Survey 2018, a2 ed., 2018.

[2] L. Schenck, "Small family farming in indonesia - a country specific outlook," Food and Agriculture Organization of the United Nations., 2018.

[3] C. S. A. of Indonesia, Vegetable Production in Indonesia, the year of 2015- 2019.

[4] M. of Agriculture, T. A. Satriani, D. Martianto, and Y. Heryatno, Food Consumption Development Directory, YulivaEditor, Ed. Ministry of Agriculture, 2019.

[5] F. Balducci, D. Fomarelli, D. Impedovo, A. Longo, and G. Pirlo, "Smart farms for a sustainable and optimized model of agriculture," in 2018 AEIT International Annual Conference, 2018, pp. 1–6.

[6] Khairunnisak, Devianti, and Mustafril, "Study of application of automatic watering equipment with drip irrigation system based on changes in groundwater level on pakcoy (brassica chinensis l.)," Unsyiah Agricultural Student Scientific Journal, vol. 2, p. 298, Aug 2017.

[7] H. Research, D. Agency, W. Setiawati, R. Murtiningsih, G. A. Sopha, and T. Handayani, Technical Instructions for Vegetable Cultivation, 2007, vol. 1.

[8] S. Jindarat and P. Wuttidittachotti, "Smart farm monitoring using raspberry pi and arduino," in 2015 International Conference on Computer, Communications, and Control Technology (I4CT), 2015, pp. 284–288.

[9] K. Sharma and R. Nandal, "A literature study on machine learning fusion with iot," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), April 2019, pp. 1440–1445.

[10] L. Aminulloh, W. T. Sesulihatien, and D. Pramadihanto, "Feature extraction of tomato growth model using greenhouse monitoring system," in 2019 International Electronics Symposium (IES), Sep. 2019, pp. 370–375.

[11] S. Gertphol, P. Chulaka, and T. Changmai, "Predictive models for lettuce quality from internet of things-based hydroponic farm," in 2018 22nd International Computer Science and Engineering Conference (ICSEC), 2018, pp. 1–5.

[12] K. S. Pratyush Reddy, Y. M. Roopa, K. Rajeev L.N., and N. S. Nandan, "Iot based smart agriculture using machine learning," in 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 130–134.

[13] S. Y. Chaganti, P. Ainapur, M. Singh, Sangamesh, and S. O. R., "Prediction based smart farming," in 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), 2019, pp. 204–209.

[14] I. Stancin and A. Jovi ˇ c, "An overview and comparison of free python libraries' for data mining and big data analysis," in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 977–982.

[15] H. Faradina, "Design and implementation of iot tolls for automation on smart farm," 2021.

[16] F. Ismail, "Design and implementation of the iot platform for monitoring pakcoy plants in the seedbed stage," 2021.

[17] D. J. P. P. S. M. P. Anuja Changude, Nikita Harpale, "A review on machine learning algorithm used for crop monitroing system in agriculture," in Internaitonal Research Journal of Engineering and Technology (IRJET), vol. 05, 2018, pp. 1468–1471.

[18] S. Ray, "A quick review of machine learning algorithms," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 35–39.

[19] M. Bramer, Principles of Data Mining, 3rd ed. Springer, 2016.

[20] K. R. Dalal, "Analyzing the role of supervised and unsupervised machine learning in iot," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 75–79.

[21] D. Fumo, "Types of machine learning algorithms you should know," Jun 2017. [Online]. Available: https://towardsdatascience.com/ types-of-machine-learning-algorithms-you-should-know-953a08248861

[22] S. Athmaja, M. Hanumanthappa, and V. Kavitha, "A survey of machine learning algorithms for big data analytics," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1–4.

[23] A. Gonfaloniery, "How to build a data set for your machine learning project," Feb 2019. [Online]. Available: https://towardsdatascience.com/ how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac

[24] J. Community, "About us," Jan 2021. [Online]. Available: https: //jupyter.org/about

[25] S. Kumar, N. Dhanda, and A. Pandey, "Data science — cosmic infoset mining, modeling and

visualization," in 2018 International Conference on Computational and Characterization Techniques in Engineering Sciences (CCTES), 2018, pp. 1–4.
[26] "Scikit-learn." [Online]. Available: https://github.com/scikit-learn
[27] "Scikit-learn notebook." [Online]. Available: https://www.kaggle.com/ notebooks?searchQuery=scikit-learn
[28] M. Eremia, C. C. Liu, and A. A. Edris, Decision Trees, 2016, pp. 819–844.
[29] S.-l. Community, "Decision tree." [Online]. Available: https://scikit-learn. org/stable/modules/tree.html
[30] T. Xie, R. Li, X. Zhang, B. Zhou, and Z. Wang, "Research on heartbeat classification algorithm based on cart decision tree," in 2019 8th International Symposium on Next Generation Electronics (ISNE), 2019, pp. 1–3.
[31] S. Kohli, “Understanding a classification report for your machine learning model,” Nov 2019. [Online]. Available: https://medium.com/@kohlishivam5522/ understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397