 doi:10.20944/preprints202201.0445.v1

Article

Internet of Things-driven Data Mining for Smart Crop Production Prediction in the Peasant Farming Domain

Luis Omar Colombo-Mendoza [1], Mario Andrés Paredes-Valverde [2], María del Pilar Salas-Zárate [3,*], Rafael Valencia-García [4]

[1] Tecnológico Nacional de México / I. T. S. Teziutlán; luis.cm@teziutlan.tecnm.mx
[2] Tecnológico Nacional de México / I. T. S. Teziutlán; mario.pv@teziutlan.tecnm.mx
[3] Tecnológico Nacional de México / I. T. S. Teziutlán; maria.sz@teziutlan.tecnm.mx
[4] Universidad de Murcia; valencia@um.es
* Correspondence: valencia@um.es

Abstract: Internet of Things (IoT) technologies can greatly benefit from machine learning techniques and Artificial Neural Networks for data mining and vice versa. In the agricultural field, this convergence could result in the development of smart farming systems suitable for use as decision support systems by peasant farmers. This work presents the design of a smart farming system for crop production, which is based on low-cost IoT sensors and popular data storage services and data analytics services on the Cloud. Moreover, a new data mining method exploiting climate data along with crop production data is proposed for the prediction of production volume from heterogeneous data sources. This method was initially validated using traditional machine learning techniques and open historical data of the northeast region of the state of Puebla, Mexico, which were collected from data sources from the National Water Commission and the Agri-food Information Service of the Mexican Government.

Keywords: data mining; predictive analytics; Internet of Things; peasant farming; smart farming system; crop production prediction.

1. Introduction

The set of techniques that allow manually and automatically extracting information that resides implicitly in data in a non-trivial way and that could be useful for various processes is known as data mining [1]. Data mining is rooted in Artificial Intelligence, especially, in Machine Learning (ML), as well as in Statistical Analysis. Through models extracted using Artificial Intelligence and Statistical Analysis techniques it is possible to solve problems that imply prediction, classification and segmentation tasks, making that large amount of data can be processed and used more efficiently [2].

Furthermore, the process of extracting information from large datasets with the aim of making estimations about future results Is known as predictive analytics. It represents an intermediate step within a broader process of data analytics known as business analytics [3]. In this context, Machine Learning can be defined as a data analysis method that automates the construction of analytical models. Its study is based on the idea that software systems can learn at least semi-autonomously from information by identifying patterns and making decisions with minimal human intervention.

Predictive analytics, along with Internet of Things (IoT) technologies, has been extensively applied to the agricultural domain in recent years [4-9]. This has enabled the development of the concepts of smart agriculture/farming and precision agriculture/farming. IoT technologies are the set of predominant and emerging Information and Communication Technologies (ICT) that are the foundation of a global infrastructure for the information society, which enables advanced services by interconnecting virtual and physical "things".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

According to the Food and Agriculture Organization of the United Nations, agriculture in Mexico represents more than an important productive sector. Beyond its contribution to the national GDP, which is barely 4%, the multiple functions of agriculture in Mexico's economic, social and environmental development indicate that its incidence is much greater than that indicator would imply.

In particular, the volume of agricultural production in the Mexican state of Puebla contributed 7 million 403 thousand 938 tons to the country's agricultural production in 2018, ranking fifteenth among the 32 states of the country. Nevertheless, this contribution implies 22.6% of the economically active population of the state of Puebla, which shows a disparity that indicates that the Puebla's farmlands are not very efficient.

Considering that the primary sector of the economy is the primary source of food and sustenance for families that live in rural communities and even in rural communities that are very far from the urban centers in Puebla, it is evident that rural development is one of the main pillars of the growth and well-being of the Puebla society.

This work seeks to contribute to the recovery of the farmland of the Mexican state of Puebla, which is one of the major purposes of its government, by proposing: (1) the architectural design of a smart peasant farming system for crop production prediction, which is based on low-cost IoT sensors and popular data storage services and data analytics services on the Cloud and (2) a new data mining method exploiting climate and crop production data sources for the prediction of the volume of production of corn grain in the Northeast region of the state of Puebla.

Figure 1 gives an overview of our research idea; it formally shows a simplified version of the workflow of the proposed system architecture, in which only the major components, and the interactions among them, are included. This figure highlights the generation of crop production predictions as output, as well as the roles of IoT-based sensors and the peasant farmer (as end user) in provisioning heterogeneous input data.

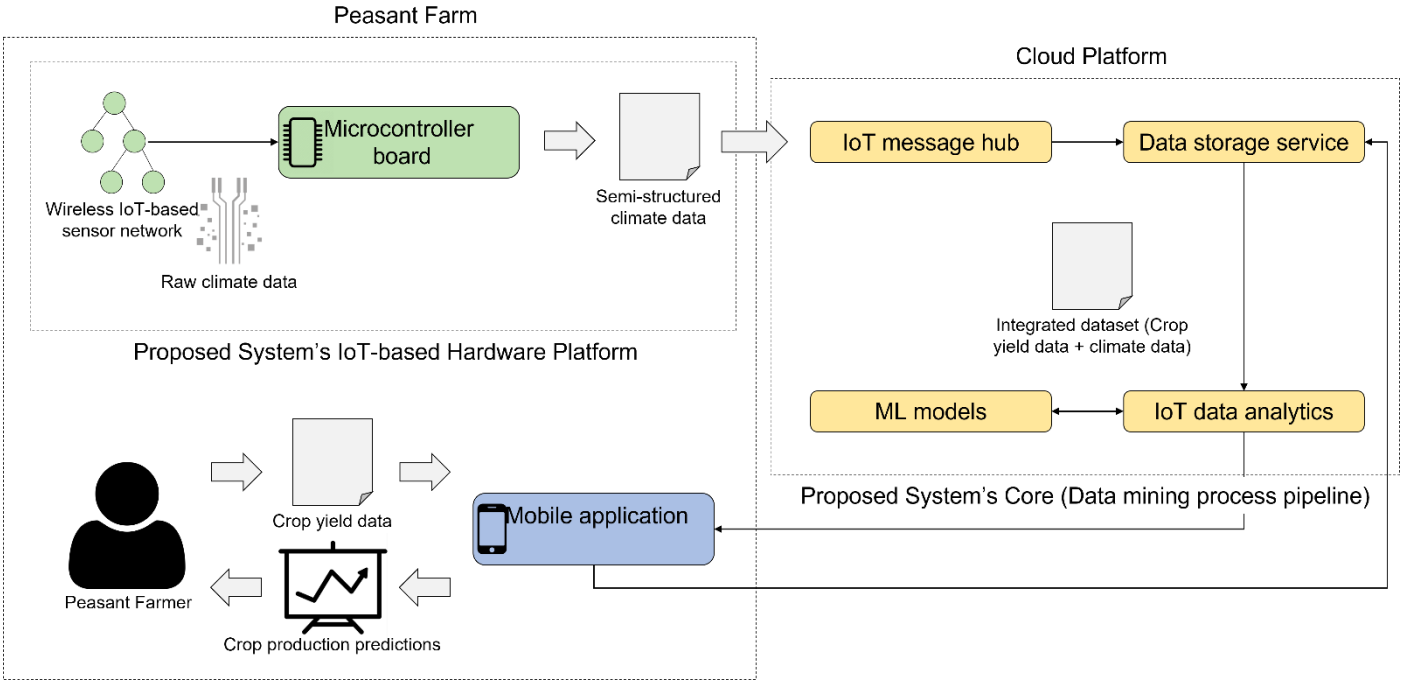

Figure 1. Research Idea Overview.

As shown in Figure 1, climate data, namely, temperature and rainfall, are automatically collected from IoT-based sensors (other climate data, namely, storm activity and fog and hail occurrence, are gathered from datasets made available by local weather stations), whereas crop yield data, namely, hectares planted, hectares harvested and actual crop production volumes, are provided by the peasant farmer though a mobile application. Furthermore, notice that the interactions among the components that represent the core

 doi:10.20944/preprints202201.0445.v1

of the system architecture: IoT message hub, data storage service, IoT data analytics service and ML models depict the main steps of the proposed data mining method.Unlike most recent related works, our work proposes a data mining process for the prediction of the volume of crop production that integrates two heterogeneous sources of crop production data and climate data as well as the architectural design of a smart farming system specially designed for the peasant farming.

The remainder of this paper is organized as follows: in the first section, some recent proposals that are relevant to our work are described and compared; in the second section, our proposal is presented in detail; in the third section, the results of an initial validation of our proposal are presented and discussed; finally, in the last section, the concluding remarks of our work are summarized and the future lines of research are outlined.

2. State of the Art

Some of the state-of-the-art proposals that are more related to our proposal are described below for comparison purposes.

A smart farming system using Internet of Things (IoT) technologies and machine learning techniques for data mining is presented in [10]. It is based on an architecture comprising the following four layers: (a) a layer consisting of IoT sensors and actuators deployed in the farming field, (b) an edge computing layer that provides integration between an IoT wireless sensor network and the cloud using IoT gateways, (c) a cloud computing layer that is intended to store and analyze data over cloud servers and (d) an end user mobile and Web-based application layer. In that work, a new prediction method is proposed as the foundation of a decision support system for crop productivity and drought prediction based on the integration of the PART classification technique and the wrapper feature selection approach.

F. Balducci et al. [11] present five cheap, practical and easy-to-implement data analysis experiments intended to increase smart farming productivity. These experiments range from forecasting of future crop harvest on complete time-series data to reconstruction of missing or wrong IoT sensors data, passing through the detection of faulty IoT sensors from the geographical clustering of source monitoring stations using the Euclidean distance metric. A variety of machine learning algorithms such as decision tree, k-nearest neighbors and linear regression, as well as a single-layer perceptron neural network, were used and compared for these purposes in conjunction with three heterogeneous datasets belonging to industry, scientific research and statistical institutions.

A precision agriculture system that seeks to reduce efforts and labor of agricultural sector personnel, which uses IoT sensors for data collection, as well as machine learning and deep learning techniques to detect damage and diseases in crops, was presented in [12]. The system is structured into the following four subsystems: (1) smart irrigation system, (2) smart fertilizer dose recommendation system, (3) crop disease detection system, and (4) crop damage prediction system. Three Convolutional Neural Network (CNN) architectures were implemented for disease prediction using multiclass image classification: ResNet50, VGG16 and DenseNet121. Regarding damage prediction, five machine learning algorithms were used: LightGBM, XGBoost, Random Forest, Decision Tree and K-Nearest Neighbors (KNN), obtaining better results with the LightGBM algorithm.

Adel et al. [13] presented an architecture of an IoT-based smart monitoring system for agriculture, which was intended to give advice to farmers to avoid and prevent the spread of the late blight disease in potatoes and tomatoes. This system consists of three different layers, namely: (a) a perception layer consisting of data acquisition nodes composed of sensors, microcontrollers and communication modules, (b) an application layer that displays all collected data to farmers through a dynamic web application and (c) a gateway layer that connects the perception layer with the application layer. Additionally, the authors implemented a prediction model using the Linear Regression technique and a classification model using a Support Vector Machine (SVM) algorithm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

With the aim of significantly contributing to the saving of freshwater used in agriculture, especially, in irrigation, an architecture of an intelligent autonomous irrigation system based on IoT technologies and machine learning techniques was proposed in [14]. In particular, the system allows predicting soil moisture using information collected from sensors deployed on the ground and weather forecast information extracted from the Internet through web services, thus helping to make effective irrigation decisions with optimal water use. A hybrid machine learning algorithm, based on a Support Vector Regression algorithm and the K-means clustering algorithm was implemented for this purpose.

Li et al. [15] presented an intelligent agriculture system for the management and control of greenhouses. The system uses different IoT devices to collect a large amount of greenhouse environmental data and an improved K-means clustering algorithm based on the maximum distance method to select relatively optimal data as reference data for the next cycle in a greenhouse. It is structured into four major layers: (1) a sensors layer, which includes a variety of sensors, video cameras and other types of data acquisition hardware, (2) a transport layer, which includes wireless communication and wired communication modules, (3) a business layer that is mainly responsible for the monitoring of the environmental data and (4) an application layer that allows interaction with the user through different web applications.

An expert system for the domain of agriculture, which is based on Artificial Intelligence techniques, specifically, on artificial neural networks, was proposed in [16]. This system helps farmers to assess land suitability for cultivation based on farming data obtained from an underlying wireless sensor network. In particular, this data is collected from different IoT-based sensors, including PH, soil moisture, salinity and electromagnetic sensors, using a Raspberry Pi Single-Board Computer (SBC), and it is then locally preprocessed and sent to the Cloud to be stored for further processing. A Multi-Layer Perceptron (MLP)-based model for classification was finally implemented that exploits data stored in the Cloud with the purpose of classifying land suitability for cultivation.

Alibabaei et al. [17] implemented Recurrent Neural Network (RNN) models, namely, Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM and Bidirectional GRU models, to estimate tomato and potato yields at the end of a season based on time series data, specifically, climate Big Data, irrigation scheduling data and soil water contents. Climate Big Data was collected by an agricultural weather station for a site in Portugal and retrieved from a government agency of the Ministries of Agriculture and the Sea. The performance of the models was compared with the performance of a Convolutional Neural Network model, a Multi-Layer Perceptron model and a Random Forest Regression model, and the results showed that Bidirectional LSTM model outperformed all alternative and baseline models in predicting tomato and potato yields.

A comparative analysis of the works described above is summarized in Table 1. This analysis comprises the following criteria of comparison: purpose, use of IoT technologies, use of data mining techniques, use of machine learning/artificial intelligence techniques, machine learning task implemented, crop studied, use of crop production data and use of climate data.

Table 1. Comparative analysis of related works.

Criterion/Work	[10]	[11]	[12]	[13]
IoT technologies	✔ (IoT sensor-based dataset collection)	✘	✘	✔ (IoT sensor-based dataset collection)
Data mining techniques	✔	✘	✔	✘

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

Machine learning/artificial intelligence techniques	✔ (PART algorithm)	✔ (Linear Regression, Single-Layer Perceptron)	✔ (Random Forest, KNN, LightGBM, RestNet50, VGG26 and DenseNet121)	✔ (SVM and Linear Regression)
Machine learning task	Classification	Time series forecasting	Classification	Prediction and classification
Crop studied	Bajra, soybean, jowar and sugarcane	Pear and apple	Tomato, potato, corn, apple and peach	Tomato and potato
Purpose	Drought and crop productivity	Crop harvest	Crop damage and disease	Crop disease
Crop production data	✔	✘ (Soil productivity data)	✘ (Crop fertilization data)	✘
Climate data	✔	✘	✘	✔
Criterion/Work	**[14]**	**[15]**	**[16]**	**[17]**
IoT technologies	✔ (IoT sensor-based soil mosture data collection)	✔ (IoT sensor-based dataset collection)	✔ (IoT sensor-based dataset collection)	✘
Data mining techniques	✘	✔	✔	✔
Machine learning techniques/artificial intelligence techniques	✔ (SVR and K-means)	✔ (K-means)	✔ (Multi-Layer Perceptron)	✔ (LSTM, GRU, Bidirectional LSTM, bidirectional GRU)
Machine learning task	Prediction and clustering	Clustering	Classification	Time series forecasting
Crop studied		Unknown	Unknown	Tomato and potato
Purpose	Soil mosture	Greenhouse environmental factors optimization	Land suitability for cultivation	Crop yield
Crop production data	✘	✘	✘	✘
Climate data	✔ (Public data available on the Internet)	✘ (Greenhouse environmental data)	✘ (Soil productivity data)	✔

As is shown in Table 1, the work by Rezk et al. [10] has most of the features considered in the comparative analysis, which means that it is the work that is most similar to our work. Nonetheless, unlike our proposal, which aims to predict volume of production of crops, in that work a classification model is proposed to classify crop productivity and drought. Additionally, we focused on predicting the volume of corn grain production in the northeast region of the Mexican state of Puebla; corn grain is one of the most widely cultivated crops in the state of Puebla, Mexico, along with coffee beans and black beans. Conversely, Rezk et al. [10] focused on classifying productivity and drought of four

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

different crops that are widely cultivated in the state of Maharashtra, India, namely, bajra, soybean, jowar and sugarcan. Furthermore, unlike most of the works analyzed, our work proposes a data mining process that integrates two heterogeneous sources of crop production data and climate data.

3. Smart Peasant Farming System and Data Mining Process

The architectural design of the smart peasant farming system, which is the salient contribution of this work, is shown in Figure 2. The components of the proposed system architecture are described in the following subsections. Additionally, the data mining process for crop production prediction is described in the context of the description of the IoT data analytics component, which is one of the major components of the system architecture.

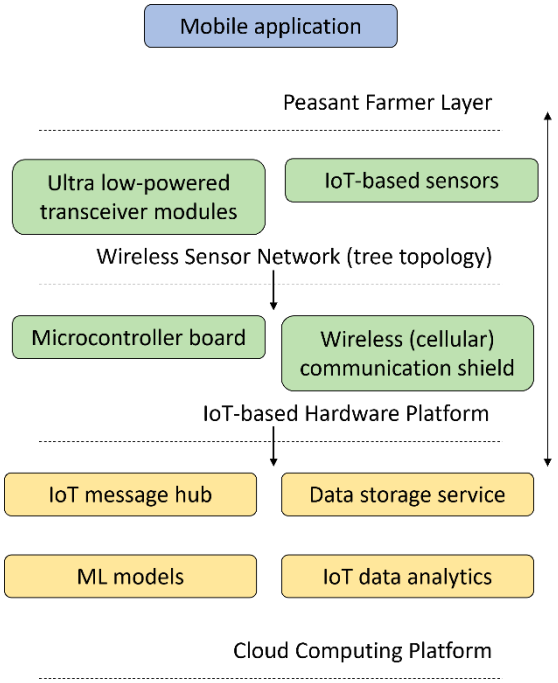

Figure 2. System Architecture.

3.1. Peasant Farming Layer

This layer basically consists in a mobile application designed for the peasant farmers to in-field register all his crop yield data; data that is commonly collected and openly published by government agencies for statistical purposes, such as the area (in hectares) sown with a crop, the harvested area (in hectares) of a crop and the yield volume (in tons) of the harvested area of a crop.

Likewise, the mobile application is intended to show the peasant farmers the results of the data mining process, i.e., the results of the crop yield prediction task on the integrated historical dataset: crop yield data + climate data. It is also intended to serve as a tool for the real-time monitoring of this data as it is remotely collected using a variety of in-field sensors. In this context, Internet of Things platforms such as Blynk, Ubidots and Arduino Cloud could be exploited to build web and mobile dashboards using user interface drag & drop editors i.e., without programming any code.

3.2. Wireless Sensor Network

The Wireless Sensor Network (WSN) represents the second layer of the architecture of our system, which is based on ultra-low powered transceiver modules designed for operation in the worldwide Industrial, Scientific and Medical (ISM) frequency band at

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

2.4GHz such as the nRF24L01+ single chip 2.4GHz transceiver [1], as well as on a variety of low-power sensors for climate and environmental monitoring. In this regard, the selection of the necessary sensors should favor low-cost sensors that are compatible with hardware and software prototyping platforms such as Arduino and NodeMCU, which are popular and relatively low-cost alternatives for the development of IoT-based systems.

The choice of this Radio Frequency (RF) wireless communication technology over other similar technologies that are equally proprietary, such as Long Range (LoRa), lies in the possibility of exploiting communication protocols that are specially designed to enable high-power data transmission and reception at lower power consumptions. In fact, in the context of peasant and family farming, high bandwidths should be prioritized over long transmission ranges as constant up links for real-time data streams are frequently deployed over narrow geographic areas.

Moreover, a typical tree topology in which one of the nodes acts as a base node and the others are central hubs or actual sensor nodes is proposed for the design of the WSN.

Unlike other network topologies for WSNs, namely, cluster topology and flat topology, tree topology has been demonstrated to save a bit more energy in data acquisition applications. On the other hand, tree topology can perform worse than cluster topology and chain topology in terms of scalability; similarly, it can perform worse than chain and flat topologies in terms of topology management (overhead) [18]. Nonetheless, a smart farming system such as the one that we pursue in this work, which is aimed at peasant and family farming, is theoretically less likely to suffer from these problems as it would be composed of no more than a few dozen sensors.

3.3. IoT-based Hardware Platform

The foundation of this layer is a general-purpose microcontroller-based development board wired to the transceiver module of the WSN acting as the root node. In this regard, some general-purpose electronic prototyping platforms such as Arduino and the family of Discovery Boards offered by STMicroelectronics have microcontroller-based development boards especially designed for the IoT, such as Arduino 33 IoT and B-L475E-IOT01A, respectively.

These microcontroller boards usually facilitate integration with IoT platforms, cloud services or mobile and web development platforms such as Blynk, Amazon Web Services, or Google Firebase, respectively; nevertheless, they tend to be relatively more expensive than general-purpose microcontroller boards, e.g., Arduino UNO.

Furthermore, one of the major components of the IoT-based Hardware Platform layer is the wireless networking module which must enable access to the Internet to communicate with the Cloud (represented by the Cloud Computing Platform layer).

Due to the characteristics of the domain of application of our system, it will not normally be located near WiFi access points, so using Wireless Local Area Networks to connect it to the Internet would not be a viable option. Therefore, we have chosen to use cellular networks as an alternative networking technology in this regard.

3.4. Cloud Computing Platform

The fourth layer of the architecture of our system is composed of four cloud computing services of four different categories: cloud messaging service, data storage service, machine learning service and IoT data analytics service.

3.4.1. IoT message hub service

In general terms, a cloud messaging service enables a channel for bi-directional communication between IoT devices/mobile applications and the Cloud. Beyond the obvious

[1] https://www.sparkfun.com/datasheets/Components/SMD/nRF24L01Pluss_Preliminary_Product_Specification_v1_0.pdf

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

need to create datastreams for all the variables that are commonly monitored by a variety of sensors in a smart farming system, bi-directional communication is crucial for a data mining-based smart farming system to be able to show back to its users the results of the data mining process that is commonly carried out in the Cloud.

3.4.2. Data storage service

A data storage cloud service typically consists of a NoSQL database (a non-relational database), either an object database or a JavaScript Object Notation (JSON)-based database. The importance of this component within the proposed architecture lies in the possibility of storing all data collected by sensors composing the WSN in a secure and fully scalable manner.

In addition, some cloud computing solutions from this category that are part of platforms aimed at developing serverless web and mobile applications such as Google Firebase and Amazon AWS Amplify allow multiple client devices to directly connect to databases through bi-directional channels. This enables real-time synchronization of data on all devices directly connected to a database in response to changes made on each of the devices, which would partially eliminate the need for a cloud messaging service. Nonetheless, there would be a need for a service that allows sending messages from the Cloud back to the connected devices because of the execution of functions hosted in the Cloud, which would be the case of the crop yield prediction task in our system.

3.4.3. IoT data analytics service

Regarding the data mining process, all data stored in the NoSQL cloud database, which is raw data, must first be preprocessed to be transformed into data that can be used by Machine Learning algorithms for the purpose of discovering knowledge. In particular, within the proposed architecture, this task is carried out by the IoT data analytics service.

Moreover, data preprocessing commonly involves data cleaning and data transformation itself. On the one hand, data cleaning means fixing or removing anomalies in data, and, in its simplest form, it is reduced to dealing with missing values, removing irrelevant values and removing duplicated values. This is crucial for an IoT sensor-based smart farming system because data captured from the physical world through sensors (in our case, climate data) tend to be noisy and unreliable.

On the other hand, data transformation typically involves data scaling and data normalization. Data scaling means fitting data within specific scales whereas data normalization implies scaling data with the intention of transforming it to be normally distributed.

In this context, it is worth noting that popular cloud platforms (Platform as a Service, PaaS) such as Microsoft Azure and Amazon AWS include services that allow automatizing common data preprocessing tasks, from data cleaning tasks through data transformation tasks. On these cloud platforms, data can also be automatically preprocessed before being stored.

Figure 3 shows a flowchart representation of the data mining process pipeline proposed in this work.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

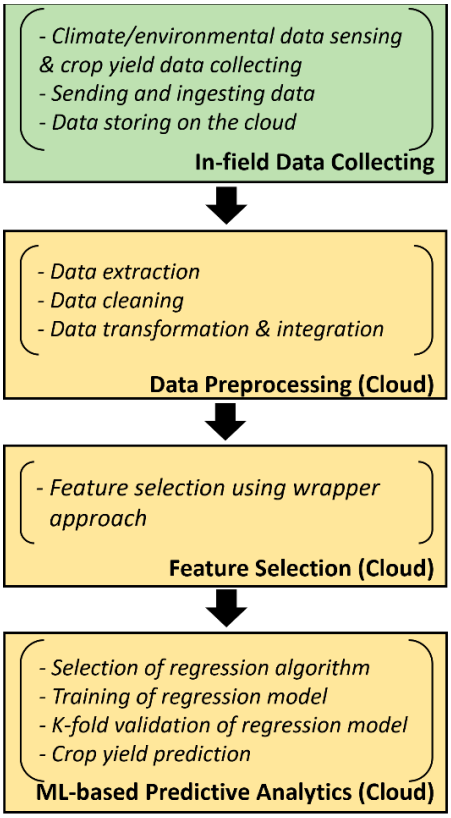

Figure 3. Data Mining Process Pipeline.

3.4.4. Machine learning service

One of the major components of the Cloud Computing Platform layer is the Machine Learning service, which allows implementing predictive algorithms and works in conjunction with the IoT Data Analytics service to enable machine learning-based predictive analytics for the purpose of predicting crop yield.

In this context, being considered a de facto computational notebook for data science, Jupyter Notebook is supported by many of the most popular cloud platforms as a rapid, iterative and interactive way of implementing machine learning algorithms for data mining, which mainly includes splitting datasets into training and validation sets as well as training and validating machine learning models. In this work, we have chosen k-fold cross-validation as the preferred method for training and validating predictive models as it is the most recommended method for machine learning model evaluation [19].

Implementing machine learning algorithms also implies selecting those that are theoretically more appropriate given the nature of the data to be processed and the type of data mining task to be carried out (in our case, prediction). In addition, it implies analyzing target data to identify those features or variables that are more relevant for use in generating machine learning models, task that is known as feature selection.

Feature selection can be carried out using two main different approaches: wrapper-based approaches and filter-based approaches. On the one hand, with the wrapper-based approaches, multiple machine learning models are evaluated using procedures that incrementally add or remove features to find the approximately optimal combination that maximizes model performance. These procedures are mostly realized by greedy search algorithms; a greedy algorithm is any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage.

On the other hand, filter-based approaches allow evaluating the relevance of the features outside of the machine learning models using statistical calculations keeping only the features that pass some criterion. Unlike filter-based approaches, wrapper-based approaches completely depend on the underlying machine learning algorithms and tend to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

be computationally intensive; they, however, usually provide the best-performing feature set for a particular type of machine learning model [20].

This component is finally responsible for making crop yield predictions by exploiting resulting machine learning models.

4. Materials & Methods

For a preliminary validation of the proposed architecture, a climate dataset was collected containing the following data observed during the period of 2003-2019 in the municipalities of Teziutlán, Tlatlauquitepec, Hueyapan, Hueytamalco, Yaonáhuac, Acateno, Atempan, Teteles de Ávila Castillo, Zaragoza, Chignautla, Ayotoxco de Guerrero, Zacapoaxtla, Cuetzalan de Progreso of the northeast region of the Mexican State of Puebla:

- Monthly average temperatures
- Days with hail events per month
- Days with fog occurrence per month
- Days with storm activity per month
- Total monthly rainfall

Similarly, a crop yield dataset was collected that contains the following data observed during the same period in the same municipalities of the Mexican State of Puebla for both corn crops:

- Hectares planted with corn
- Hectares harvested for corn
- Production volumes (in tons) of the harvested areas of corn

Figure 4 shows a map of the previously mentioned municipalities of the Mexican State of Puebla.

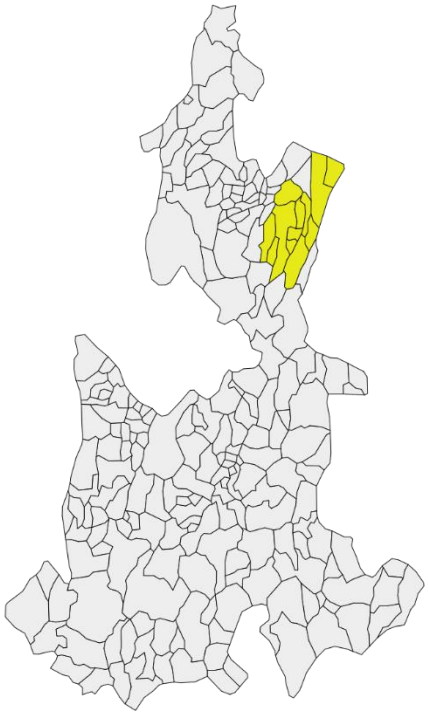

Figure 4. Selected municipalities of the Mexican State of Puebla.

These datasets were manually collected from the website of Mexico's National Water Commission [2] and the website of Mexico's Agri-food Information Service [3], respectively.

[2] https://smn.conagua.gob.mx/es/informacion-climatologica-por-estado?estado=pue

[3] http://infosiap.siap.gob.mx/gobmx/datosAbiertos.php

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

The latter publishes annualized data on agricultural production at national, state and municipal levels as open data available in the form of spreadsheet files, whereas the former publishes a variety of data collected by the country's weather stations as annualized data by state in plain text format.

Table 2 summarizes the locations of the weather stations of the northeast region of the Mexican State of Puebla used as sources of climate data in this study.

Table 2. Weather stations.

Weather Station Name	Municipality	Location
"Teziutlán"	Teziutlán	Latitude: 19°49'49" N. Longitude: 097°21'00" W. Altitude: 1,818.0 MASL
"Oyameles"	Tlatlauquitepec	Latitude: 19°42'51" N. Longitude: 097°32'51" W. Altitude: 2,670.0 MASL
"Las Margaritas"	Hueytamalco	Latitude: 19°59'14" N. Longitude: 097°17'14" W. Altitude: 2,422.0 MASL
"San José Acateno"	Acateno	Latitude: 20°08'24" N. Longitude: 097°12'04" W. Altitude: 144.0 MASL
"Zaragoza"	Zaragoza	Latitude: 19°47'10" N. Longitude: 097°33'10" W. Altitude: 2493.0 MASL
"Los Humeros (CFE)"	Chignautla	Latitude: 19°40'45" N. Longitude: 097°24'22" W. Altitude: 2,862.0 MASL
"Ayotoxco de Guerrero"	Ayotoxco de Guerrero	Latitude: 20°05'43" N. Longitude: 097°25'43" W. Altitude: 237.0 MASL
"Zacapoaxtla (SMN)"	Zacapoaxtla	Latitude: 19°52'18" N. Longitude: 097°35'18" W. Altitude: 1,828.0 MASL
"Cuetzalan de Progreso"	Cuetzalan de Progreso	Latitude: 20°02'20" N. Longitude: 097°31'20" W. Altitude: 756.0 MASL

As shown in Table 2, for most of the municipalities of our interest (9 out of 13), there is one local weather station providing data for the period of reference (2003-2019), which we used as climate data source in this study.

Moreover, Table 3 summarizes the number of instances, features and missing values in each dataset. Notice that these statistics correspond to the unprocessed datasets. Details of handling missing data and reducing features (selecting features) are given in the following subsections.

Table 3. Unprocessed datasets' statistics.

Dataset	Number of Instances	Number of Features	Number of Missing Values
Climate dataset	221	62	1209
Crop yield dataset	442	6	0

The following section describes the variables included in the second dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

4.1. Crop Yield Dataset Description

Among the main variables included in the crop yield dataset are those described in Table 4. This dataset includes other variables used as reference: year,production cycle name and municipality name, which are not shown in Table 4. The year and municipality name variables are also included in the climate dataset.

Table 4. Variables in crop yield dataset.

Variable	Description	Unit of Measurement
Total cropped area	The total area planted with a crop	Hectares
Total harvested area	The total area harvested for a crop	Hectares
Production volume	The harvested production of a crop	Tons

4.2. System Construction

We partially implemented the proposed architecture to preliminarily validate it. First, we used NodeMCU as the microcontroller-based development board and the SIM 900 GSM/GPRS shield as the wireless networking module for the realization of the IoT-based Hardware Platform layer.

Second, we used Amazon's AWS IoT Core, AWS IoT Analytics and S3 services to realize the IoT Message Hub, IoT Data Analytics and Data Storage Service components of the Cloud Computing Platform layer of the proposed architecture. Likewise, we used Jupyter Notebook documents to implement machine learning algorithms for predictive analytics in Python, being able to run the resulting models on AWS IoT Analytics thanks to the integration of this cloud service with the Jupyter Notebook data science tool.

4.3. Data Preprocessing on the Cloud

The datasets collected were ingested into Amazon S3 buckets, then this data was sent from the Amazon S3 service to the AWS IoT Analytics service for data preprocessing purposes.

In particular, the climate and crop yield datasets were clean and transformed separately and then integrated into a single dataset. The integrated dataset comprised approximately 400 samples or observations.

Regarding data cleaning, the crop yield dataset required minimum treatment. On the contrary, the climate dataset required deeper treatment due to missing data (see Table 3). Any missing value of the variables representing monthly average temperatures, total monthly rainfalls, days with hail events in a month, days with fog occurrence in a month and days with storm activity in a month were calculated as the average of all the values registered for the corresponding month for all the years included in the dataset. In this context, notice that there are novel proposals for data mining methods that inherently deal with missing values [21].

Regarding data transformation, average temperatures per production cycle (spring-summer and autumn-winter production cycles), as well as total rainfall per production cycle, days with hail events per production cycle, days with fog activity per production cycle and days with storm activity per production cycle were accordingly calculated from the variables representing monthly average temperatures, total monthly rainfalls, days with hail events in a month, days with fog occurrence in a month and days with storm activity in a month in the case of the climate dataset. As a result, the climate dataset was restructured as shown in Table 5.

Table 5. Variables in climate dataset.

Variable	Description	Unit of Measurement

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

Average temperature_spring	Average temperature during spring-summer production cycle.	Degrees Celcius
Average temperature_autumn	Average temperature during autumn-winter production cycle.	
Days with hail activity_spring	Days with hail activity during spring-summer production cycle.	Days
Days with hail activity_autumn	Days with hail activity during autumn-winter production cycle.	
Days with fog occurrence_spring	Days with fog occurrence during spring-summer production cycle.	
Days with fog occurrence_autumn	Days with fog occurrence during autumn-winter production cycle.	
Days with storm activity_spring	Days with storm activity during spring-summer production cycle.	
Days with storm activity_autumn	Days with storm activity during autumn-winter production cycle.	
Total rainfall_spring	Total rainfall during spring-summer production cycle.	Millimeters
Total rainfall_autumn	Total rainfall during autumn-winter production cycle.	

For the integration of the crop yield and climate datasets into a single dataset, we implemented a database-style joining approach by which new observations were generated by joining the observations from the latter with those from the former using the year and municipality name variables as indexes (see Figure 5). The rationale behind this is that crop production data must be interpreted in the context of the data about the weather conditions of the cropping areas.

From this perspective, the importance of the climate data in this work lies in the use that we made of it to enrich the crop yield data.

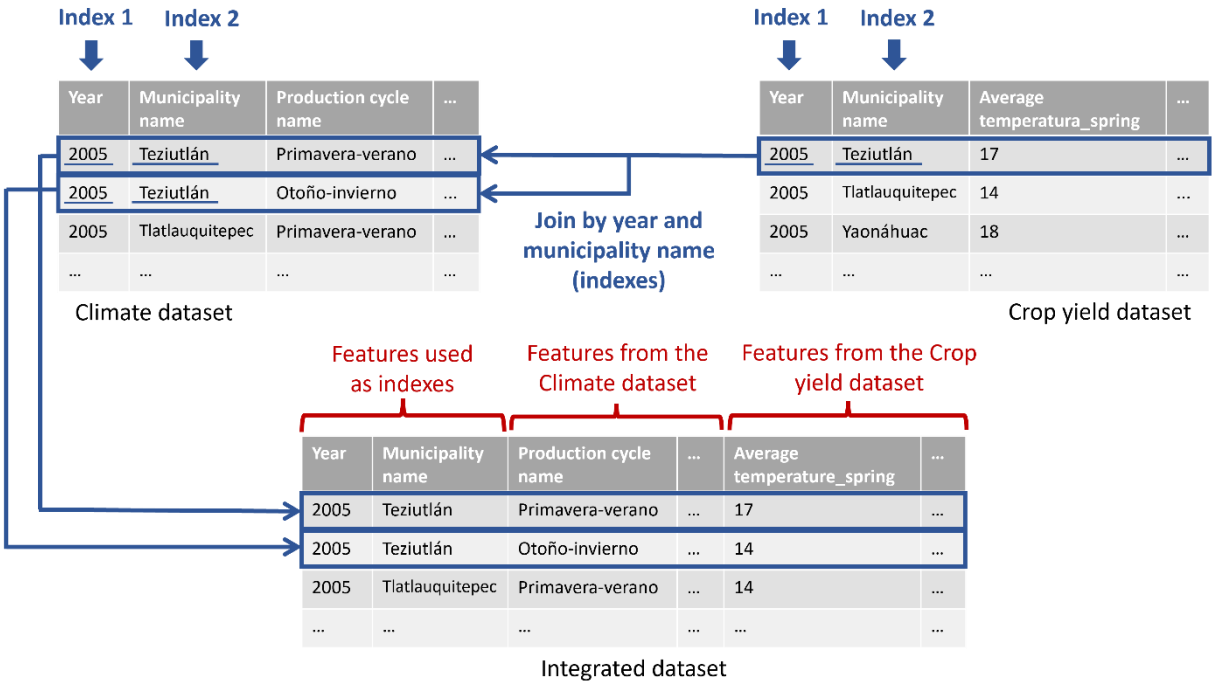

Figure 5. Database-style joining approach for dataset integration.

In addition, categorical variables in the integrated dataset, namely, production cycle name and municipality name were transformed into numerical variables using dummy variables.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

Finally, range of data in the integrated dataset was rescaled into a [0, 1] range, i.e., it was normalized.

4.4. Feature Selection and ML-based Predictive Analytics on the Cloud

For feature selection, we preferred a wrapper-based approach to a filter-based approach. Because of this choice, we performed the feature selection step in conjunction with the Machine Learning-based Data Analytics step of the proposed data mining method.

In particular, we implemented the Recursive Feature Elimination (RFE) method [22] to perform feature selection on the integrated dataset. RFE is an instance of the Backward Feature Elimination method, which consists in an iterative process that implies training a classifier or regressor, computing the ranking criterion for all features involved and removing the feature with the smallest ranking criterion.

Furthermore, we used the Linear Regression and K-Nearest Neighbors (KNN) Regression machine learning algorithms in the core of the RFE algorithm as we selected them for the implementation of machine learning-based predictive analytics in this work. In regard to the latter, we experimentally set the number of neighbors (k) at 5.

We carried out the following iterative procedure to select the optimal *k* value:

1. Initialize a random *k* value between 2 and N-1 (where N is the number of samples in our dataset).
2. Train the prediction model using the selected *k* value. Use the trained model to make predictions for the data in the dataset reserved for validation.
3. Calculate the Root Mean Square Error (RMSE) for the predictions computed.
4. Repeat the process selecting a different *k* value.

We finally created a plot between the RMSE values and the *k* values to select the *k* value having the minimum error rate, which was 5 (*k*=5).

Linear Regression and KNN Regression are, respectively, easy and simple parametric and non-parametric machine learning algorithms [23, 24], and they were judged as theoretically appropriate for the nature of both the problem that we faced and the data that was available.

We sought to significantly reduce the number of features to employ to build prediction models using these algorithms; therefore, we experimentally set the threshold of relevant features at nine by carrying out the following iterative procedure:.

1. Initialize a random value between 2 and F-1 (where F is the total number of features in our dataset) for the number of features to be selected (*f*).
2. Perform Recursive Feature Elimination (RFE) to select the *f* most relevant features from the set of features in our dataset.
3. Train the prediction model using the selected *f* most relevant features. Use the trained model to make predictions for the data in the dataset reserved for validation.
4. Calculate the Root Mean Square Error (RMSE) for the predictions computed.
5. Repeat the process selecting a different *f* value.

We finally created a plot between the RMSE values and the *f* values to graphically select the f value having the minimum error rate, which was 9 (*f*=9).

For both prediction models, the procedure resulted in the selection of the following nine features from our integrated dataset: total cropped area, total harvested area, year, total rainfall_spring, total rainfall_autumn, average temperature_spring, average temperature_autumn, days with hail activity_spring and days with hail activity_autumn.In particular, we trained and validated the prediction models using the k-fold cross-validation method, for which we set the number of folds (k) at 5 [19]. As a result, the integrated dataset was split into 5 consecutive folds from which 1 was used once for test while the 4 remaining folds were used once to train the models.

5. Results

Table 6 shows the results of the calculation of the default performance metric for prediction models in the scikit-learn library, in this case, for both the Linear Regression model

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

and the K-Nearest Neighbors Regression model, namely, Coefficient of Determination (R^2) [25]. This metric provides a measure of how well the model is likely to predict unseen samples through the proportion of variance that is explained by the independent variables in the model [26]. In particular, this table shows the means of the R^2 scores computed for both prediction models in each step of the 5-fold cross-validation, as well as the standard deviations of these R^2 scores.

Table 6. Coefficient of Determination (R^2) scores.

Model	Coefficient of Determination (R^2) Scores	
	Mean	Standard Deviation
Linear Regression Model	0.756	+/-0.005
KNN Regression Model	0.944	+/-0.001

We must be careful in judging the high mean R^2 scores, because according to some statistical tests that we carried out, there is collinearity between some of the predictor variables in our integrated dataset. We decided not to address this issue in this study because collinearity does not tend to influence the ability of a prediction model to predict new observations [26-27] and the goal of this study was to make accurate predictions. Nevertheless, in future work, we need to analyze to what extent these high R^2 scores are an indication of the collinearity problems existing in our integrated dataset.

Additionally, the standard deviations of the R^2 scores suggest that there is a very little variation in the performance of our prediction models when using different subsets of training data.

Furthermore, to support the R^2 scores obtained, we created estimated-by-observed plots for the learned Linear Regression Model and the learned KNN Regression Model. Furthermore, we performed a graphical residual analysis on the results of the models to assess the assumptions of the regression models. In particular, we created a residuals vs. fits plot and a histogram of residuals for each of the models. Notice that a residual represents the vertical distance between an observed data point and its estimated value.

Figure 6shows the estimated-by-observed plots, whereas the residuals vs. fits plots are shown in Figure 7 and the histograms of residuals are shown in Figure 8.

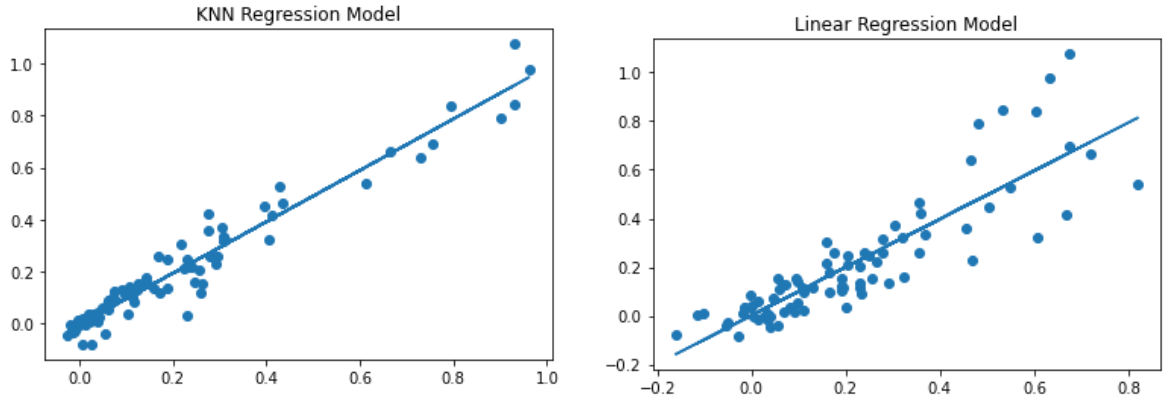

Figure 6. Predicted-by-observed plots.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

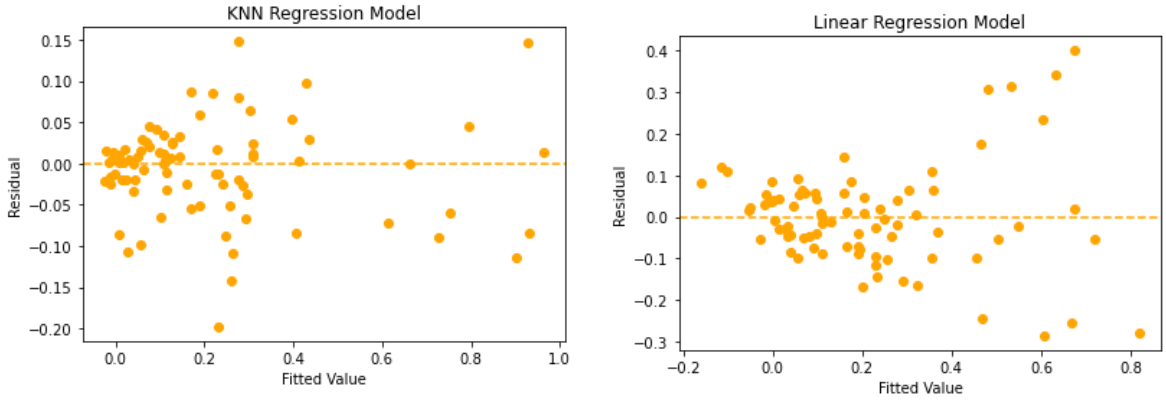

Figure 7. Residuals vs. fits plots.

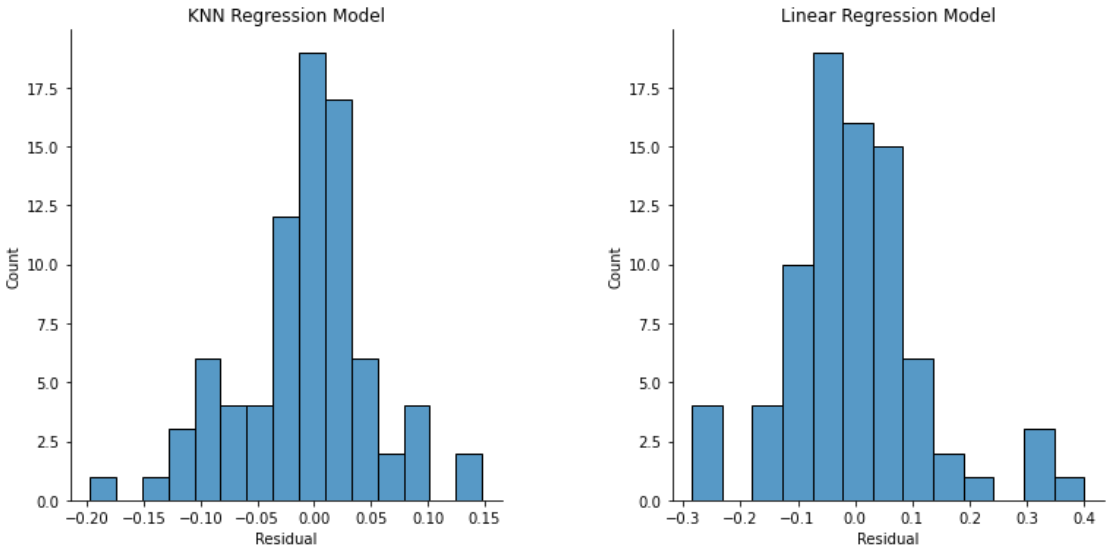

Figure 8. Histograms of residuals.

As shown in Figure 6, the estimated values of the KNN Regression Model are more strongly correlated with the observed values than the estimated values of the Linear Regression Model. This is a clear indication of how accurate each model is with respect to the other, and it supports the R^2 scores obtained, which can be interpreted as follows for the Linear Regression Model and the KNN Regression Model, respectively:

- 75.6% of the variation in response y (production volume) is accounted for by the variation in the set of predictors X.
- 94.4% of the variation in response y (production volume) is accounted for by the variation in the set of predictors X.

Recall that the set of predictors X resulting from the feature selection process comprised the following predictors (features): total cropped area, total harvested area, year, total rainfall_spring, total rainfall_autumn, average temperature_spring, average temperature_autumn, days with hail activity_spring and days with hail activity_autumn.

Furthermore, as shown in Figure 7, the residuals of both the Linear Regression Model and the KNN Regression Model are randomly scattered around the residual = 0 line, which indicates that the assumption of linear relationship is reasonable; nonetheless, the average of the residuals remains closer to 0 in the case of the KNN Regression Model than in the case of the Linear Regression Model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0445.v1

In addition, as shown in this figure, the variation of the residuals appears to be roughly constant at every level of the fitted values for the KNN Regression Model, which is evidence that the assumption of constant variance is not violated. The residuals vs. fits plot of the Linear Regression Model shows approximately eight data points that can be judged as outliers as they do not follow the general trend of the rest of the data. Therefore, we should be careful in stating that the assumption of constant variances is also not violated in the case of the Linear Regression Model; we should perform tests of equality of variances to check it in future work.

Finally, the bell-shaped appearance of the histograms of the residuals of the Linear Regression Model and the KNN Regression Model, which are shown in Figure 8, is a clear indication that the assumption of normality is reasonable.

Moreover, we computed error rates for our final prediction models. In particular, we selected the Root Mean Square Error (RMSE) metric, which is a risk metric that corresponds to the expected value of the square root of the quadratic error or loss (see Table 7). RMSE is a very commonly used general-purpose metric for numerical predictions [28].

Table 7. Root Mean Square Error (RMSE) rates.

Model	Root Mean Square Error (RMSE) Rate
Linear Regression Model	0.122
KNN Regression Model	0.058

Unlike R^2 scores, which should be close to 1, RMSE rates should be close to 0. As shown in Table 7, the RMSE rates for the Linear Regression Model and the KNN Regression Model are 0.122 and 0.058, respectively, which are certainly close to 0. In addition, the RMSE rate of the KNN Regression Model is closer to 0 than the RMSE rate of the Linear Regression Model. This finding is in correspondence with the finding of the previous performance analysis, which was carried out based on the R^2 metric, and it represents a strong indication of how well each prediction model perform with respect to the other.

6. Discussion

Judging by the results of the evaluation performed, it is feasible to use data mining techniques to integrate heterogeneous crop production and climate data and to exploit it using traditional machine learning techniques to reliable predict the volume of production of the crops (t).

One of the major challenges that we faced in the integration of crop production data and climate data was the difference in data granularity: crop production data was available as annualized data whereas climate data was available as monthly data. This led us to perform simple temporal aggregation on the climate data, which intuitively implies the loss of some information as the number of observations is reduced. Notice that this information loss could naturally influence prediction performance negatively.

A linear relationship between the production volume variable and other crop production and climate variables such as total cropped area (t), total harvested area (t), average temperature (°C) and total rainfall (mm) was proven to exist using corn grain production data and climate data from the northeast region of the Mexican state of Puebla.

Crop production predictions can be considered a starting point for the farmers to be able to make effective decisions in a timely manner based on reliable findings made before harvesting and even just after planting at the beginning of a production cycle. In fact, the potential benefits that crop production predictions could have in planning crop production cycles should be assessed in future work. It is clear, however, that for the peasant farming families in the northeast region of the State of Puebla, Mexico, the possibility of accessing low-cost technology that allows them to maximize crop production is of great importance, considering that, in many cases, agricultural production directly represents their primary source of food.

In future work, a smart peasant farming system should also be constructed based on the architecture proposed in this work to be able to compile a dataset from crop production and climate data collected using the set of in-field IoT-based sensors and the mobile application conceived for that purpose. This will allow us to carry out a validation of our proposal as a whole. Alternatively, regarding climate data, we should explore the use of satellite imagery as a possible data source. In fact, the other big challenge that we faced in this study was the shortage of climate data due to the scarcity of local weather stations across the selected municipalities of the Mexican State of Puebla.

Additionally, we will study the suitability of using other traditional supervised machine learning algorithms and artificial intelligence techniques such as artificial neural networks to build more accurate prediction models. In this context, we plan to explore a hybrid supervised/unsupervised machine learning approach in which clustering algorithms (unsupervised machine learning) are used to automatically label our data and machine learning algorithms for prediction are used to make predictions based on the labeled data [29-30].

Regarding feature selection, we will study the feasibility of solving our prediction problem using regression techniques that directly reduce the set of predictive variables to the smaller set of uncorrelated variables, such as Partial Least Squares (PLS) Regression and Principal Component Regression.

Furthermore, we will address the problem of predicting the volume of production of other crops that are popular in the northeast region of the state of Puebla, Mexico, such as coffee beans and black beans, including data that allows capturing climate variability such as the minimum and maximum temperatures.

Finally, we plan to study the feasibility of integrating Semantic Web technologies for knowledge representation and reasoning and recommendation techniques based on these technologies to the proposed architecture to improve the data mining process [31-32].

Author Contributions: All authors contributed equally to the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tecnológico Nacional de México, grant number 337533 (11079).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. I. H. Witten, E. Frank, M. A. Hall, y C. J. Pal, “Chapter 1 - What’s it all about?”, in Data Mining (Fourth Edition), 4th ed., I. H. Witten, E. Frank, M. A. Hall, y C. J. Pal, Eds. Morgan Kaufmann, 2017, pp. 3–41. doi: 10.1016/B978-0-12-804291-5.00001-5.
2. M. J. Zaki y W. M. Jr, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd ed. Cambridge University Press, 2020.
3. D. Delen, Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners, 2nd ed. Hoboken: Pearson FT Press, 2020.
4. N. Dlodlo y J. Kalezhi, “The internet of things in agriculture for sustainable rural development”, in 2015 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), may 2015, pp. 13–18. doi: 10.1109/ETNCC.2015.7184801.
5. A. Tzounis, N. Katsoulas, T. Bartzanas, y C. Kittas, “Internet of Things in agriculture, recent advances and future challenges”, Biosystems Engineering, vol. 164, pp. 31–48, dec. 2017, doi: 10.1016/j.biosystemseng.2017.09.007.
6. P. P. Ray, “Internet of things for smart agriculture: Technologies, practices and future direction”, Journal of Ambient Intelligence and Smart Environments, vol. 9, no. 4, pp. 395–420, jan. 2017, doi: 10.3233/AIS-170440.
7. X. Shi et al., “State-of-the-Art Internet of Things in Protected Agriculture”, Sensors, vol. 19, no. 8, Art. no. 8, jan. 2019, doi: 10.3390/s19081833.
8. L. Colizzi et al., “Chapter 1 - Introduction to agricultural IoT”, in Agricultural Internet of Things and Decision Support for Precision Smart Farming, A. Castrignanò, G. Buttafuoco, R. Khosla, A. M. Mouazen, D. Moshou, y O. Naud, Eds. Academic Press, 2020, pp. 1–33. doi: 10.1016/B978-0-12-818373-1.00001-9.
9. Y. He, Q. Zhang, y P. Nie, “Introduction of Agricultural IoT”, in Agricultural Internet of Things: Technologies and Applications, Y. He, P. Nie, Q. Zhang, y F. Liu, Eds. Cham: Springer International Publishing, 2021, pp. 1–21. doi: 10.1007/978-3-030-65702-4_1.

10. N. G. Rezk, E. E.-D. Hemdan, A.-F. Attia, A. El-Sayed, y M. A. El-Rashidy, "An efficient IoT based smart farming system using machine learning algorithms", Multimed Tools Appl, vol. 80, no. 1, pp. 773–797, jan. 2021, doi: 10.1007/s11042-020-09740-6.
11. F. Balducci, D. Impedovo, y G. Pirlo, "Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement", Machines, vol. 6, no. 3, Art. no. 3, sep. 2018, doi: 10.3390/machines6030038.
12. S. Garg, P. Pundir, H. Jindal, H. Saini, y S. Garg, "Towards a Multimodal System for Precision Agriculture using IoT and Machine Learning", arXiv:2107.04895 [cs], jul. 2021.
13. A. A. Araby et al., "Smart IoT Monitoring System for Agriculture with Predictive Analysis", in 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), may 2019, pp. 1–4. doi: 10.1109/MOCAST.2019.8741794.
14. A. Goap, D. Sharma, A. K. Shukla, y C. Rama Krishna, "An IoT based smart irrigation management system using Machine learning and open source technologies", Computers and Electronics in Agriculture, vol. 155, pp. 41–49, dec. 2018, doi: 10.1016/j.compag.2018.09.040.
15. C. Li y B. Niu, "Design of smart agriculture based on big data and Internet of things", International Journal of Distributed Sensor Networks, vol. 16, no. 5, p. 1550147720917065, may 2020, doi: 10.1177/1550147720917065.
16. D. R. Vincent, N. Deepa, D. Elavarasan, K. Srinivasan, S. H. Chauhdary, y C. Iwendi, "Sensors Driven AI-Based Agriculture Recommendation Model for Assessing Land Suitability", Sensors, vol. 19, no. 17, Art. no. 17, jan. 2019, doi: 10.3390/s19173667.
17. K. Alibabaei, P. D. Gaspar, y T. M. Lima, "Crop Yield Estimation Using Deep Learning Based on Climate Big Data and Irrigation Scheduling", Energies, vol. 14, no. 11, Art. no. 11, jan. 2021, doi: 10.3390/en14113004.
18. Q. Mamun, "A Qualitative Comparison of Different Logical Topologies for Wireless Sensor Networks", Sensors (Basel), vol. 12, no. 11, pp. 14887–14913, nov. 2012, doi: 10.3390/s121114887.
19. S. Ozdemir, "12. Beyond the Essentials", in Principles of Data Science: Learn the techniques and math you need to start making sense of your data: Mathematical techniques and theory to succeed in data-driven industries, 1st ed., Birmingham, UK: Packt Publishing, 2016.
20. M. M. Mafarja y S. Mirjalili, "Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection", Soft Comput, vol. 23, no. 15, pp. 6249–6265, aug. 2019, doi: 10.1007/s00500-018-3282-y.
21. D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, 'Clustering mixed numerical and categorical data with missing values', Information Sciences, vol. 571, pp. 418–442, Sep. 2021, doi: 10.1016/j.ins.2021.04.076.
22. I. Guyon, J. Weston, S. Barnhill, y V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, vol. 46, no. 1, pp. 389–422, jan. 2002, doi: 10.1023/A:1012487302797.
23. B. Boehmke y B. Greenwell, "Chapter 4 Linear Regression", in A Machine Learning Algorithmic Deep Dive Using R, 1st ed., Boca Ratón, FL: Chapman and Hall / CRC, 2020.
24. A. Bartosik y H. Whittingham, "Chapter 7 - Evaluating safety and toxicity", in The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry, S. K. Ashenden, Ed. Academic Press, 2021, pp. 119–137. doi: 10.1016/B978-0-12-820045-2.00008-8.
25. S. Glantz, B. Slinker, y T. Neilands, "Chapter Three: Regression with Two or More Independent Variables", in Primer of Applied Regression & Analysis of Variance, Third Edition, 3rd ed., New York: McGraw Hill / Medical, 2016.
26. M. Kutner, C. Nachtsheim, J. Neter, y W. Li, "Chapter 2 Inferences in Regression and Correlation Analysis", in Applied Linear Statistical Models, 5th ed., New York: McGraw-Hill / Irwin, 2004.
27. J. Frost, 'Chapter 9. Checking Asssumptions and Fixing Problems', in Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models, 1st ed., Statistics by Jim Publishing, 2020.
28. S. P. Neill and M. R. Hashemi, 'Chapter 8 - Ocean Modelling for Resource Characterization', in Fundamentals of Ocean Renewable Energy, S. P. Neill and M. R. Hashemi, Eds. Academic Press, 2018, pp. 193–235. doi: 10.1016/B978-0-12-810448-4.00008-2.
29. D.-T. Dinh, T. Fujinami, and V.-N. Huynh, 'Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient', in Knowledge and Systems Sciences, Singapore, 2019, pp. 1–17. doi: 10.1007/978-981-15-1209-4_1.
30. D. Astolfi and R. Pandit, 'Multivariate Wind Turbine Power Curve Model Based on Data Clustering and Polynomial LASSO Regression', Applied Sciences, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/app12010072.
31. R. Valencia-García, J. M. Ruiz-Sánchez, P. J. Vivancos-Vicente, J. T. Fernández-Breis, and R. Martínez-Béjar, "An incremental approach for discovering medical knowledge from texts", Expert Systems with Applications, vol. 26, no. 3, pp. 291–299, Apr. 2004, doi: 10.1016/j.eswa.2003.09.001.
32. F. García-Sánchez, R. Colomo-Palacios, and R. Valencia-García, "A social-semantic recommender system for advertisements", Information Processing & Management, vol. 57, no. 2, p. 102153, Mar. 2020, doi: 10.1016/j.ipm.2019.102153.