VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2024, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 21, Number 2, April-June 2024 pp:192-212

A Model for Wheat Yield Prediction to Reduce the Effect of Climate Change Using Support Vector Regression

Abdul Sami Kharal [1], **Shahid Ali Mahar** [1*], **Muhammad Imran Mushtaque** [1], **Aurangzeb Magsi** [1], **Javed Ahmed Mahar** [1]

[1]Institute of Computer Science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan

Keywords: *Wheat Yield Prediction, Support Vector Regression, Outliers Detection, Simple Linear Regression, Multiple Linear Regression*

Journal Info:
Submitted:
May 25, 2024
Accepted:
June 20 , 2024
Published:
June 30, 2024

Abstract Recent changes in the climatic conditions have significantly threatened the food security globally. Increasing in temperature adversely affected different crops in Pakistan particularly Wheat crop. Mostly farmer's crop wheat in District Khairpur but yield is not predicted yet. Therefore, famers are unable to estimate the effects of climate changes. This research work introduces a novel framework for the development of wheat yield prediction model using Support Vector Regression. The model incorporates four predictor variables: temperature, rainfall, humidity and pH value of soil. The essential wheat yield data obtained from official departments, websites, and scholarly publications. Five datasets are created from the gathered data in order evaluate the suggested wheat prediction model. For the creation of dataset, some preprocessing operations such as handling missing values and outlier's detection are applied to the collected raw data. Experiments performed using simple linear and multiple linear regression models. By dividing the dataset in 70% and 30%, model training and testing performed respectively. The conducted research illustrated that multiple linear regression model provide desired outcomes

***Correspondence author email address:** shahid.mahar@salu.edu.pk
DOI: 10.21015/vtse.v12i2.1855

1 Introduction

With billions of people depending on agriculture as their primary means of food and a living, it is the most strategically important sector globally [1]. Pakistan's economy is mostly dependent on agriculture, which makes up a significant proportion of its GDP. In Pakistan, agriculture produces over 45% of all jobs and contributes about 23% in the country's GDP. Roughly 60% of those living in rural areas are exclusively dependent on agriculture for their standard of living [2]. Agricultural land and water are the most important natural resources of Pakistan, which has an abundance of them across a range of ecosystems and climate types [3]. Pakistan, a developing nation with limited GDP, depends heavily on its agriculture industry to feed its expanding population, which has

increased by more than four times. Persistent and secure nutrition supply can only be guaranteed by taking comprehensive steps including forming foundation, executing maintainable practices and remaining up to date with innovation [4].

Traditional approaches used for estimating crop production take too much time and human effort [5]. Therefore, modern approaches with the help of technology provide better results for prediction in agriculture along with insightful information on factors affecting crop production. In this regard, regression analysis is a very popular and efficient method for accurately evaluating multidimensional data, highlighting the relationship between important variables [6, 7]. Regression analysis considers a number of characteristics for this task such as yearly rainfall, temperature, soil pH value and others as independent variables, while crop yield is considered as a dependent variable in these frameworks. Even though regression analysis is a very powerful tool for yield prediction tasks, it still has not gained much consideration for predicting wheat production in Pakistan. Adequate forecasting regarding the yield is very important for agrarian policymakers, based on which different crop management strategies and resource allocation is performed. Moreover, there are various problems encountered with the techniques used for wheat yield prediction including poor prediction methods, less importance given to key factors affecting wheat yield, and many more.

A statistical technique called regression analysis is employed to ascertain the correlation among a single dependent and one or more independent variables. Wheat has been considered for this research because of its vital significance in the region's agriculture, especially considering Pakistan's vast agricultural environment [8, 9]. This study develops a novel framework to achieve the task of wheat yield prediction for Khairpur District using regression analysis. Regression analysis is a popular method to deal with complex data that uses a mathematical formulation to explain the link between a number of predictor factors (also known as independent variables) and a major result variable (also known as dependent variable). Local government organizations provided the primary data used in this investigation. It is essential to carry out a thorough investigation and introduce an effective regression analysis strategy that takes into account a number of weather factors, agricultural and soil properties techniques. This will improve the precision and accuracy of the wheat crop yield prediction model. This study aims to increase the accuracy and applicability of crop prediction simulation approaches specifically in Pakistan, which would result in more efficient crop management methods and superior agricultural output. The major research contribution which shows the innovative aspect of this study lies in its comprehensive approach to analyzing complex data including detailed analysis and incorporation of various parameters into the regression model.

2 Literature Review

Computerized agricultural methods have been the subject of numerous research articles with a focus on crop production using different techniques. Since agriculture is the key factor in the food production process, its significance for preserving the food supply increases with the global population growth [10]. In the modern era, advanced and progressive farming overcomes the old practices. It combines several information technology domains like robotics, cloud computing, machine learning, and more. It is crucial to understand how environmental changes impact the agriculture sector because agricultural production has a significant impact on the global economy [11]. There is also a need to control variables that greatly affect agricultural productivity, such as crop genetics, farming practices, fertilizer application, and others.

Predicting crop yields with sufficient accuracy is very important for promoting the decision-making process regarding crop reproduction [12]. An ensemble learning approach focused on the application of multisensor data fusion using unmanned aerial vehicles (UAVs) in wheat yield production. Various research studies highlight the implications of agriculture for Pakistan's economy [13]. According to Pakistan's Economic Survey for 2011–2012, 21% of the nation's GDP and 45% of its workforce are employed in agribusiness.

Irrigation is an essential part of Pakistan's farming cycle. 84% of the 22.05 million hectares of agricultural land are under irrigation, with the remainder being rainfed, also known as barani. More than 90% of the freshwater resources in the country are used for irrigated agriculture. A model has been generated to cope with the difficulties farmers have, especially those pertaining to losses brought on by a lack of understanding of how to cultivate in various soil and weather circumstances [14]. The model used artificial intelligence (AI) algorithms to give farmers information about which crops are best to cultivate based on a variety of characteristics. Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNN), and Support Vector Machines (SVM) are some of the machine learning and deep learning approaches used in its design. These methods analyze several factors that influence cultivation results, making it easier to anticipate which crops are best for a specific area. The model's main objective is to lower farmers' expenses by assisting them in selecting the best crops based on a thorough evaluation of predictive parameters.

Currently, a large number of Larkana rice farmers are unable to predict their crop yields with any degree of accuracy [15]. The primary goal of the study was to apply Regression Analysis (RA) techniques for the analysis and prediction of rice yields. Data on rice was obtained from the Agriculture Statistics Department web portal in Islamabad, which included information on yield, production, and area under cultivation. The connection among the dependent factor (yield, Y) and the independent factor (area under cultivation, AUC) for rice crops was investigated in this study using RA. The dependent and independent variables showed a positive, direct, and significant association, as demonstrated by the R^2 value. The significance of agriculture to the economy of Pakistan was emphasized by the scholars. Evaluation of opinions, rule-based learning, machine learning, genetic algorithms, random forest, classification tree, and social network analysis were among the statistical and data mining approaches researchers covered. The study highlighted the importance of crop yield prediction for food production as well as its noteworthy economic impact, especially in a nation like Pakistan that depends heavily on agriculture [16].

Using Hyper3DNetReg, the job of crop yield estimation was attempted as a two-dimensional regression problem, allowing production values for small geographic areas within a field to be simultaneously predicted. Numerous solutions based on artificial neural networks have proven to perform better in anticipating yields when compared to conventional methods. Additionally, the authors recommended using Feedforward Neural Networks in an equivalent AdaBoost algorithm for impaired learners. Furthermore, an accurate wheat production model has been created using artificial neural networks [17]. The 71-year dataset, which spans 1948 to 2018, was broken down into testing and training subsets. The model's performance was optimized through the application of hyperparameter tuning. The comparative analysis of published papers is given in Table1.

3 Methodology

In order to accomplish the necessary prediction results, the approach of this research project is divided into four primary components, each of which is further broken down into sub-modules. Data gathering is the first step in the process, when official sources are consulted to obtain information on a variety of variables, including temperature, rainfall, humidity, and soil pH. The variable selection process, which comes after data collection, identifies two categories of variables: the goal variable (wheat yield) and predictor variables (temperature, rainfall, humidity, and soil pH).

Following the selection of the variables, preparation steps are taken with the data. To make sure the data is prepared for modeling, this phase involves data purification and statistical analysis approaches. Regression equations are developed, including support vector regression, basic linear regression, and multiple linear regression models, following preprocessing. To make predictions, such algorithms are tested and trained on both single and multiple predictive variables. Lastly, a comparative study is done using the outcomes of the models that were generated. A visual illustration of the process is given

Ref. No.	Year	Technique Name	Remarks
[6]	2022	Multivariate Adaptive Regression Splines	The obtained results of yield prediction were compared among the multiple regions of India.
[10]	2022	Long short-term memory neural network and random forest (LSTM-RF)	The obtained result presented that the LSTM-RF is an efficient model.
[11]	2019	Multiple Regression Analysis	The percentage deviation of error is approximately ±30% in most of the years.
[12]	2024	Machine Learning Methods	The results presented that according to performance the XGBoost algorithm is outstanding in multi-sensor data synthesis.
[17]	2019	Artificial Neural Network (ANN)	The obtained results disclose that the most important features in wheat production are usual tendencies, momentum, and volatility.
[20]	2021	Machine Learning Models	The results are reliable with the consideration of meteorological impacts on wheat yield, suggesting the usefulness of explainable machine learning in climatological crop yield prediction in the rainy season.
[21]	2024	Support Vector Machines (SVM) and Random Forest (RF)	The SVM and RF algorithms effectively classified winter wheat fields. The SVM performance was outstanding compared to RF.

Table 1. Comparative Analysis of Wheat Yield Prediction

in Figure1.

4 Regression Models Development

Predicting a target variable as a function of multiple independent variables is possible with regression analysis, which is mainly used for forecasting. In convinced situations, it reveals the associations involving these dependent and independent values of the preferred variables as well. Following are the steps to create a regression model to forecast wheat crop yield:

Algorithm: Methods for predicting crop production through regression

Input: An exploratory dataset containing soil, crop, and meteorological data

Output: Crop yield prediction for the testing dataset.

Method

Step-1: Gather, prepare, and arrange the information: To construct a model, raw data is insufficient on its own. To produce results that are usable, the data must be gathered, sorted in accordance with the requirements, and processed effectively. Further pertinent information may be added during this procedure.

Step-2: Divide the dataset into subsets for testing and training: Two separate sets of data must be created from the data. The model is trained on the majority of the examples to anticipate the output using the training set, which normally consists of about 70% of the data. The testing set, which is comprised of the remaining data, is used to assess the system's efficiency.

Step-3: Apply regression on training sets: The algorithm's complexities should be synchronized with the problem's details, and the framework may be specified correspondingly. Throughout processing, the design and composition can be updated as required.

Step-4: For each model, determine the R^2 and residual parameters.

Step-5: To recalculate the residual and R^2 statis-

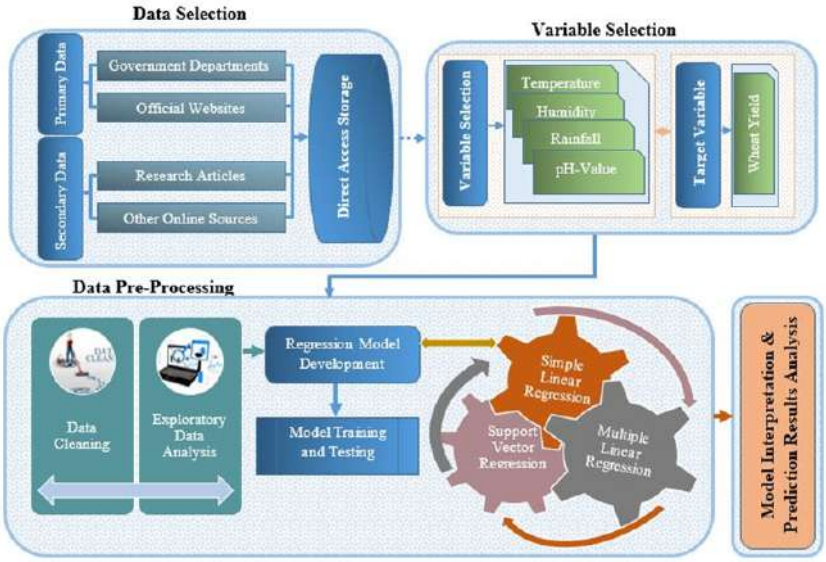


Figure 1. Proposed Framework for Wheat Yield Prediction

tics, apply the trained regression algorithm to the testing dataset. Comparing these values with other regression models is recommended. The model that provides the greatest R^2 value is deemed the most successful for crop yield prediction.

A regression model generates an equation that justifies the correlation among an independent and one or more response variables. Modeling a linear equation to the collected data allows one to generate a regression model that simulates the correlation between various explanatory variables and an outcome variable. The dependent variable in a regression model comprises a quantitative or rate/ratio element that the model seeks to explicate. The dependent variable (y) has an equivalent value for every value of the external variable (x). The regression model is created in this study to forecast wheat yield, where crop yield is the dependent, target, and response variable. The explanatory variables also include a variety of numerical or rate/ratio fields. Regression models are constructed with these independent variables to help justify the variable which is treated as dependent. For example, independent variables for a regression model predicting crop yield could be soil pH, rainfall, temperature, and humidity. These mentioned variables are mainly focused on during the experiments of wheat yield prediction using multiple statistical models of regression.

5 Data Collection

Pakistan is considered an agricultural country because of its physiography; nonetheless, province Sindh is the 2nd prolific for wheat production. The cultivated lands filled with rich silt soils formed by the Indus River are home to the wheat producing regions of Lower Sindh. Temperature affects the wheat development process in this area; shorter-duration types that harvest in 100–120 days are more common in the southeast. The data of wheat crops taken from the District Khairpur for experiments, particularly from the Taluka Gambat, is a necessary portion of this study since the Agricultural Statistics Department in Islamabad (https://www.pbs.gov.pk/) show thorough statistical information on wheat crops. District, provincial, and national levels of this data are accessible. In particular, manually compiled wheat data (in million tons) covering the years 2015–2022 for Taluka Gambat is provided in Table 2. The required data is collected with six variables i.e., Temperature Maximum (TM), Temperature Minimum (TMI), Humidity (H), Rainfall (R), pH Value of Soil (pHVS) and Crop Yield (CY).

Year	TMI (°C)	TM (°C)	H	R (MM)	pHVS	CY (Million Tons)
2015	40.82	26.73	24.2	158	46.34	362.53
2016	41.52	27.08	26.03	86.6	36.2	368.21
2017	42.15	27.32	30.78	30.1	32.26	378.16
2018	41.98	26.63	26.9	39	38.21	361.98
2019	41.47	27.57	25.37	19	45.93	385.02
2020	40.59	27.61	24.29	96.6	48.12	369.41
2021	41.83	26.45	31.08	152	52.73	372.74
2022	40.38	27.96		698	52.33	362.56

Table 2. Data Collection using Six Variables

6 Experiments

Compiling the datasets from many sources is essential to doing the research. A thorough collection of data must be preprocessed in order to obtain the necessary outcomes before being fed into the regression model.

6.1 Data Preprocessing Operations

Data preparation is an essential step in the context of a regression analysis-based wheat yield forecasting approach. The raw data needs to be cleaned and converted before it can be evaluated. The gathered data is subjected to the four preprocessing procedures listed below:

6.1.1 Handling Missing Values

Addressing missing values in the dataset to approximate the original values as closely as possible is a critical aspect of the data preprocessing. The reliability of recovering missing values largely depends on the compatibility between the dataset type and the preferred analytical method [18]. This study introduces a novel method for handling missing values in datasets that exhibit clustering properties.

The core idea of this approach is to estimate missing values by combining regression and clustering techniques [19]. To construct a preliminary dataset D, initially, the regression method is used to estimate the missing values in the specific dataset D. Subsequently, validation and clustering analysis on D is performed to identify the most effective clustering of all data records. Since records within the same cluster share significant similarities, the missing values for D in each cluster are re-estimated using regression, but only within those specific clusters. By incorporating cluster properties into the regression analysis, the accuracy of estimated missing values is enhanced. This approach integrates regression and clustering methods to address the missing value problem. The 2-dimensional (X,Y) dataset illustrated in Figure 2 shows a well-constructed linear regression model with a high coefficient of determination (R^2), indicating good data locality and model fit.

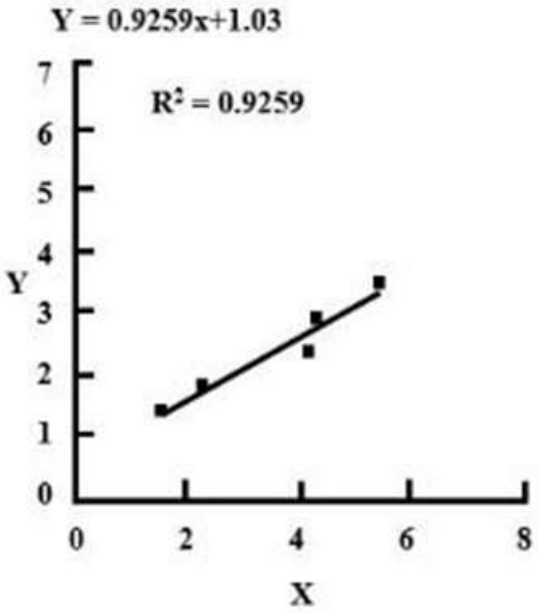


Figure 2. Dataset with Good Regression Model

By employing the Regression and Clustering (RC) method to handle missing values in a given dataset *D*, the RC method can be broken down into three key

steps as follows:

1. The dataset D is divided into two subsets, D_c and D_m, where D_m comprises all the data records with missing values, and D_c is the subset without missing values ($D_c = D - D_m$). Initially, the regression method is applied to D_c, which serves as the basis for recovering the missing values in D_m. This results in an interim dataset D', where the missing values have been imputed.
2. A clustering analysis is then conducted on D' to determine the optimal clustering configuration. Suppose that k clusters are formed, denoted as $C_1, C_2, \ldots, C_k$, ensuring that the total number of records across all clusters $\sum |C_i| = |D|$.
3. Within each cluster C_i, regression analysis is carried out for all records R_j that belong to D_m and C_i. The regression base for this prediction consists of the set $\{R_c \mid R_j \in D_c \wedge R_j \in C_i\}$, ensuring that missing values within each cluster are accurately estimated using the records from the same cluster.

6.1.2 Outlier Detection and Treatment

For quantitative data analyses to be meaningful, it's crucial that the unit of measurement remains consistent throughout the dataset. Identifying anomalies, or outliers, is a key aspect of data analysis and cleaning. Data points that substantially deviate from the main part of the dataset are called outliers. Statistical outlier detection methods form the basis of data cleaning techniques in this field. These methods identify values that significantly deviate from what is expected based on the rest of the data. The two kinds of datasets that have been the subject of the most research on context-aware outliers are temporal (time-based) and spatial datasets [20, 21].

6.1.3 Exploratory Data Analysis (EDA)

Wheat is a Rabi crop, and in the Khairpur District, its sowing begins in November, with harvesting taking place in April and May each year. Therefore, the fluctuating values of independent variables throughout the year are not suitable for processing. Consequently, for further preprocessing, the average values of temperature, monthly rainfall, relative humidity, and soil pH from November to May of each year are needed. Furthermore, the gathered data is subjected to important Exploratory Data Analysis (EDA) approach in order to provide an initial comprehension of the data dispersion. This process includes computing fundamental statistics like standard deviation, median, mean, maximum, and minimum values for variables such as wheat yield and additional pertinent predictors. Exploratory Data Analysis (EDA) entails thoroughly examining datasets to summarize their essential features using a combination of statistical computations and visual representations [22].

6.1.4 Exponential Moving Average

To examine a variable's average throughout duration, consider exponential moving averages [22]. The average of a particular set of data points can be calculated using this metric. When a new data point is added, the existing ones in the dataset are all removed. A moving average helps analysts smooth fluctuations and create a more consistent curve from a dataset that may have variations. Unlike a simple moving average, which averages data over a fixed period, the exponential moving average places more emphasis on recent data points, enhancing their impact on the overall average.

With the help of this method, the idea of weighting is introduced. Contemporary data points get more weight than older ones, which has a greater impact on the moving average computation. Equation 6.1.4 outlines the formula used to calculate the exponential moving average:

$$\mathrm{EMA}_{\mathrm{Current}} = \left(\mathrm{Value}_{\mathrm{Current}} \times \left(\frac{\mathrm{Smoothing}}{1 + \mathrm{days}}\right)\right) + \mathrm{EMA}_{\mathrm{Yesterday}} \times \left(1 - \left(\frac{\mathrm{Smoothing}}{1+\mathrm{days}}\right)\right) \quad (1)$$

The exponential moving average calculates a mean at specific intervals by assigning different weights to various data points, with the most recent values having the highest weight. This approach is beneficial for both long-term forecasting and short-term fluctuation analysis. Figure 3 illustrates the monthly temperature data over the past seven years. Similarly, exponential

moving averages for annual yield of wheat crop, humidity, and rainfall are computed using Equation 6.1.4.

6.2 Support Vector Regression

After completing the preprocessing steps, the next phase involves prediction using regression models. Both linear and non-linear regression models are employed for this purpose. Initially, a non-linear regression model, specifically the support vector regression (SVR) model, is utilized for predicting crop yield. Since training points away from a given margin are not taken into account by the model's cost operation, a subset of the training data is used while building a support vector classification system. Correspondingly, a subclass of data is used for constructing the SVR model. The SVR is formulated as the minimization of a specific objective function. Equation 2 outlines the various parameters involved in these equations.

$$\frac{1}{2}\|\mathbf{W}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2}$$

With constrains, when y = wx + b,

$$y_i - \mathbf{W}\cdot\mathbf{x}_i - b \leq \epsilon + \xi_i$$
$$\mathbf{W}\cdot\mathbf{x}_i + b - y_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

where **W** represents the weights, C is a regularization parameter, ξ_i and ξ_i^* are slack variables, ϵ represents the margin, and $\mathbf{x}_i$ and y_i are the input variables and output (target) values, respectively.

The developed dataset is divided into 70% for training and 30% for testing purposes for measuring the required prediction results. The prediction is conducted using each independent variable individually. Initially, the model is applied using temperature as the sole predictor variable, followed by rainfall, then humidity, and finally the pH value of the soil, each considered independently. Figure 4 depicts the experiments carried out using the SVR model for predicting wheat yield based on temperature as the independent variable

Figure4 demonstrated how well the model forecasts wheat yield when compared to temperature data, using training. Subsequently, the model had been developed to forecast wheat production in relation to other predictor factors, including temperature, rainfall, humidity, and soil pH. SVR prediction using many predictor variables is shown in Figure5

6.3 Linear Regression Model

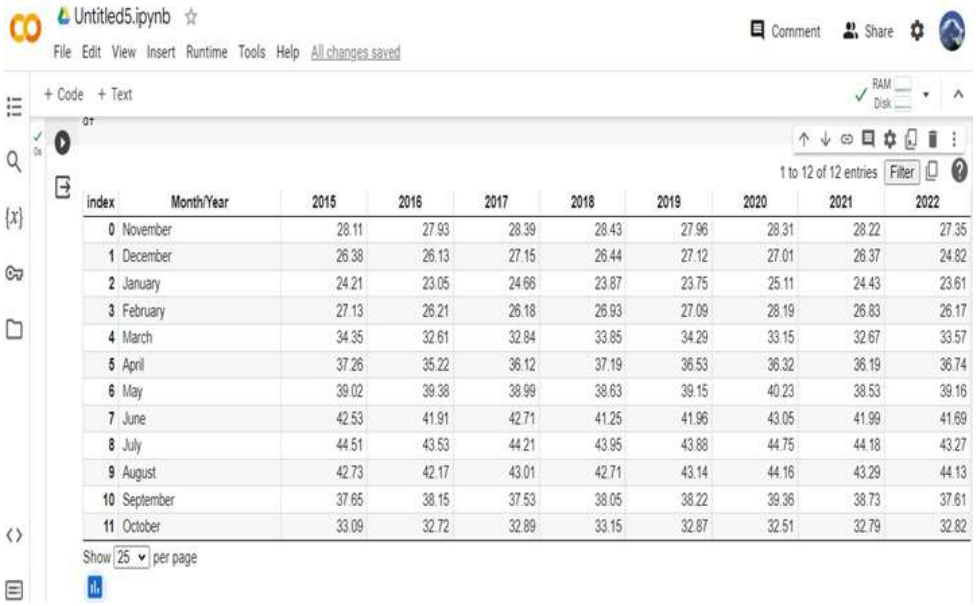
In order to create a regression model, apparent data is integrated with a linear equation to determine the affiliation between the multiple variables which are used as independent and experimented along with the targeted variable. The response variable, also known as the dependent variable, can take the form of a numerical data. Each value of the independent variable (x) corresponds to a specific value of the dependent variable (y).

The dependent variable is the crop yield, and scholars use a regression model to forecast the yearly production of wheat crops. In addition, independent variables can be incorporated into the regression model to provide an explanation for the dependent variable. According to model data, the results for the Durbin-Watson test, regression formula, R2, modified R2, p-value, residual standard error, and F statistic are described.

The accuracy of a model and the precision of its predictions for the dependent variable can be assessed through its outputs and statistical measures. The predictive capability of a model relies on the dataset it utilizes. A regression model generates a result dataset containing input fields along with estimated values, residuals, and standardized residuals. It is important to apply the Ordinary Least Squares (OLS) method to fit a linear equation to collected data, which models the relationship between independent and dependent variable.

Results like projected values and the remainder are essential for testing the OLS concepts. The values for each independent variable and the regression formula are used to determine the approximate values. The predicted values and the actual values optimally match.

The basic form of a regression equation involves



Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

1 to 12 of 12 entries Filter

index	Month/Year	2015	2016	2017	2018	2019	2020	2021	2022
0	November	28.11	27.93	28.39	28.43	27.96	28.31	28.22	27.35
1	December	26.38	26.13	27.15	26.44	27.12	27.01	26.37	24.82
2	January	24.21	23.05	24.66	23.87	23.75	25.11	24.43	23.61
3	February	27.13	26.21	26.18	26.93	27.09	28.19	26.83	26.17
4	March	34.35	32.61	32.84	33.85	34.29	33.15	32.67	33.57
5	April	37.26	35.22	36.12	37.19	36.53	36.32	36.19	36.74
6	May	39.02	39.38	38.99	38.63	39.15	40.23	38.53	39.16
7	June	42.53	41.91	42.71	41.25	41.96	43.05	41.99	41.69
8	July	44.51	43.53	44.21	43.95	43.88	44.75	44.18	43.27
9	August	42.73	42.17	43.01	42.71	43.14	44.16	43.29	44.13
10	September	37.65	38.15	37.53	38.05	38.22	39.36	38.73	37.61
11	October	33.09	32.72	32.89	33.15	32.87	32.51	32.79	32.82

Show 25 per page

Figure 3. Dataset development using Monthly Average Temperature

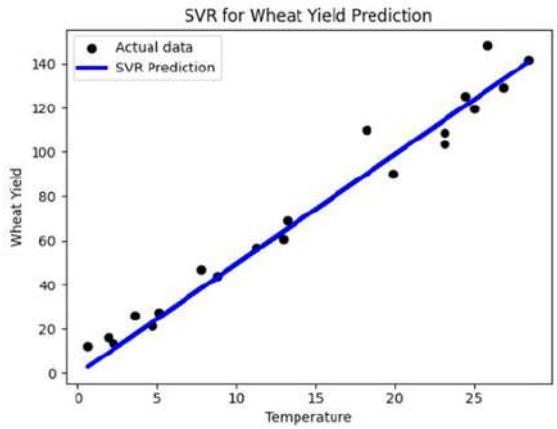


Figure 4. SVR Model Prediction with Temperature Data

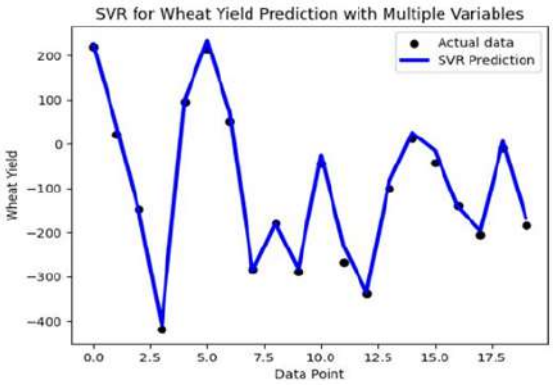


Figure 5. SVR Model Prediction with Multiple Predictor Data Variables

using estimated values alongside observed values to compute residuals. In regression analysis, residuals represent the disparities between observed values in a dataset and the values predicted by the regression equation. In a predictive model where crop yield is the label variable, temperature, rainfall, and pH value serve as training features. Equation 3 for this simple regression model is outlined as follows:

$$Y = m \cdot X + c \tag{3}$$

In this context, 'Y' represents the dependent variable, also referred to as the label, while 'X' is the independent variable, also known as the training feature. A model illustrating the relationship between these two variables is shown in Figure 6 Two data points are plotted: the first at coordinates (x_1, y_1) and the second at (x_2, y_2). Here, 'Y' is the dependent variable that relies on the x-axis, with 'X' being the independent variable. The slope of the line, denoted as 'm', is calculated using the formula $m = \frac{y_2 - y_1}{x_2 - x_1}$, and 'c' is the line's intercept. This model is employed to predict crop yield using both simple and multiple linear regression methods. For a basic linear regression model, the dependent variable (crop yield) is estimated by considering each individual

learning parameter in consequently, i.e., rainfall, temperature, humidity, and soil pH value in turn.

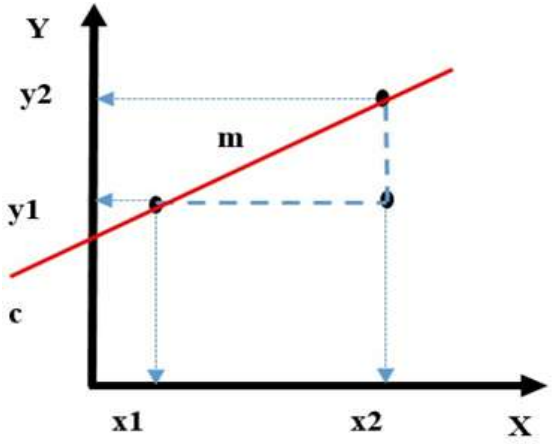


Figure 6. Simple Linear Regression Model

6.3.1 Correlation

Especially in arid land conditions, temperature, rainfall, humidity, and soil quality can all have a major effect on crop output. These factors primarily determine agricultural yield variability. Numerous factors can interact to significantly affect the productivity of wheat crops. It is expected that there will be a positive correlation between rainfall and wheat yield, as water availability is essential for all stages of growth, from germination to grain development. Wheat growth is influenced by temperature, and within a perfect range, there is typically an upward correlation among temperature and wheat yield. However, extreme temperatures can adversely affect yield. Soil pH levels also play a critical role by impacting nutrient availability. Given that soil pH stimulates nutrient uptake, there may be a positive association among wheat yield and pH levels within an ideal range. The approach of finding the correlation between the selected variables is mandatory for achieving the required outcomes.

6.3.2 Correlation between Temperature and Rainfall

The relationship among rainfall and temperature is vital in determining wheat crop yield. The nature and strength of this correlation can greatly affect crop yield and productivity. Wheat requires ideal temperature conditions for development and its growth. These early phases of development benefit from a positive association between moderate temperatures and precipitation. Similarly, maintaining a balance between temperature and water availability is critical during the vegetative phase. Moderate temperatures combined with adequate rainfall positively influence healthy vegetative growth. In contrast, prolonged periods of high temperatures and inadequate rainfall can lead to heat stress during the growing season. This stress reduces the number of seeds per spike and negatively impacts wheat yield.

A relationship between temperature and precipitation is determined using the Pearson correlation coefficient (PCC) algorithm. The Pearson correlation coefficient is a statistical metric used to measure the strength and direction of a linear relationship between two variables. It quantifies the degree to which temperature and rainfall are linearly related. By applying PCC, the linear relationship between average monthly temperatures and rainfall is evaluated each year, gaining insights into how these climatic factors interact and influence each other over time. Variable relationships can be positive (high values of X, like rainfall, are typically correlated with large quantities of Y, like temperature), negative (high values of X are typically correlated with small quantities of Y, and vice versa), or neither. PCC is reflected in the following statistical formula:

$$r_{XY} = \frac{\sum(XY) - \frac{\sum X \sum Y}{N}}{\sqrt{[\sum X^2 - \frac{(\sum X)^2}{N}][\sum Y^2 - \frac{(\sum Y)^2}{N}]}} \quad (4)$$

Where N is the total number of data points, X is the x-score (temperature), and Y is the y-score (rainfall).

To determine the correlation between various variables, the dataset is divided into different sections. Figure7 illustrates an example of the input dataset used to identify the correlation between temperature and rainfall.

The correlation matrix is shown in Figure8, along with a simulation depicting the correlation operation that has been carried out to determine the relationship between rainfall and temperature with regard to crop yield.

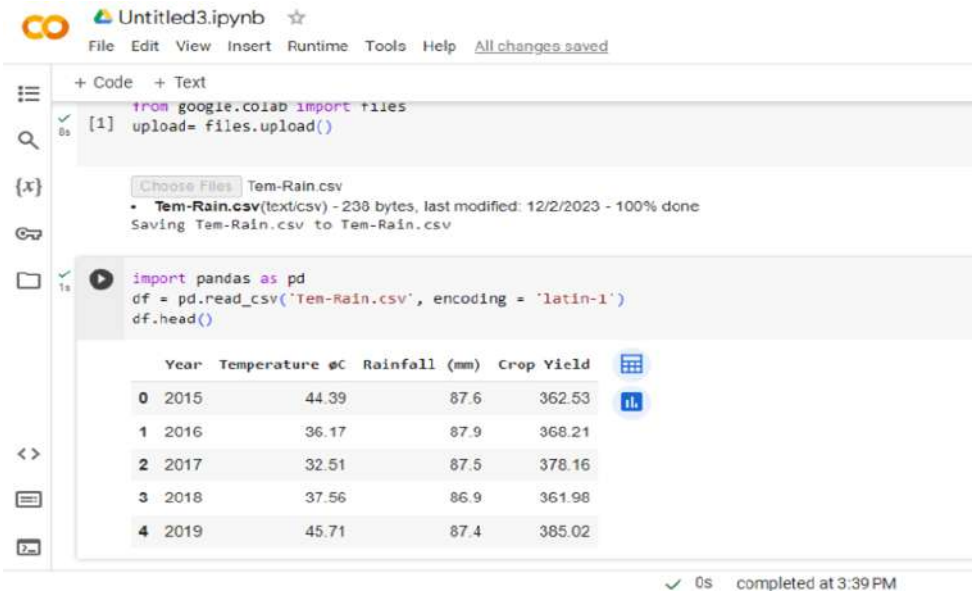


Figure 7. Selected Input Values to Execute the Operation of Correlation

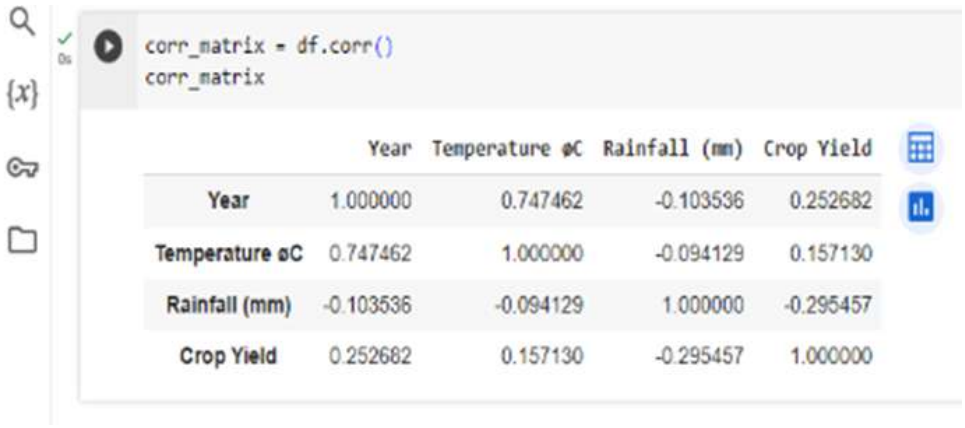


Figure 8. Correlation Operation Performed to Identify the Relations

Figure9 depicts a Heatmap illustrating the correlation analysis between temperature and rainfall. The darker regions on the Heatmap indicate areas with a very low correlation coefficient, whereas the brighter regions represent a very strong correlation coefficient between the variable.

6.3.3 Correlation between Temperature and Humidity

Humidity has an influence on wheat crop output in addition to rainfall. The rate of perspiration can be affected by humidity levels, which might impact the water imbalance of the plant. In order to calculate the extreme heat that crops endure, the Temperature-Humidity Index (THI) combines measurements of humidity as well as temperature.

Elevated THI values, which arise from high temperatures combined with high humidity, can adversely affect wheat yield by disrupting plant metabolism and inhibiting grain filling. On the other hand, low humidity combined with high temperatures can cause moisture stress. Figure10 and Figure11 show the experiments conducted to investigate the relationship between temperature and humidity.

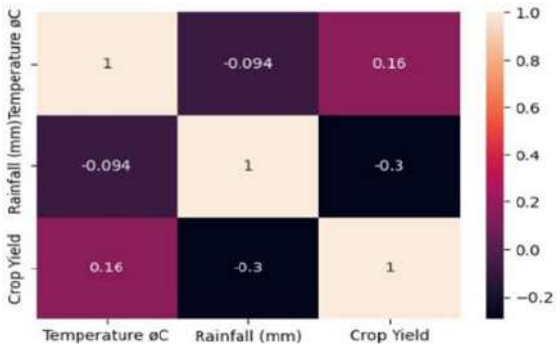


Figure 9. Explanation of Heatmap for Operation of Correlation

Figure 11 illustrates the inverse relationship between the variables "Temperature" and "Humidity." A negative value for humidity indicates that an increase in temperature leads to a decrease in humidity, which ultimately has a detrimental effect on yield production. Figure12 provides a Heatmap representation of the correlation analysis conducted between rainfall and temperature.

6.3.4 Correlation between Rainfall and Humidity

There might be seasonal and regional variations in the relationship between rainfall and humidity. Elevated humidity levels may indicate impending rain in certain regions, but not in others. Local topography can also affect the correlation between humidity and rainfall. Seasonal changes can impact this relationship as well. For optimal growth, wheat crops require sufficient water. Crops receive water from rainfall, but relative humidity can affect water availability with method like transpiration.

The excessive humidity can provide an ideal environment for some pests and illnesses that could harm wheat harvests. Excessive rainfall, particularly in areas with poor drainage, can result in waterlogging, which adversely affects crop health. During the critical grain-filling stage, sufficient moisture is crucial for wheat yield. However, too much rainfall during this period can cause lodging, which negatively impacts the yield. The experiments conducted to determine the relationship among the factors temperature and humidity are displayed in Figures13and Figure14.

Figures 14 indicate that the variables 'Humidity' and 'Rainfall' are positively correlated, meaning an increase in rainfall results in higher humidity levels. However, the relationship between humidity/rainfall and crop yield is inverse; higher humidity or rainfall tends to lead to decreased crop yield. The Heatmap depicting the association between humidity and rainfall in Figure15.

6.4 Simple Linear Regression

Following the SVR model prediction, the next step in this research methodology involves employing a simple linear regression model. This phase focuses on predicting crop yield using each independent variable individually. Initially, experiments are conducted using Temperature as the sole variable for prediction. Temperature is used as an explanatory variable in experimentation for estimating wheat production, as shown in Figure16.

Humidity, chosen as the second variable for the simple linear regression model, has a complex impact on wheat growth and yield. Its influence is intertwined with other environmental factors like temperature and soil moisture. The interactions among humidity and these variables are pivotal in shaping the overall effect on wheat production. The research findings of predicting wheat production using humidity are shown in Figure17.,

Rainfall's complicated impacts on wheat output can vary depending on a number of variables, such as the quantity, timing, and dispersion of the precipitation. Hence, predictions of wheat yield are also conducted using rainfall as an independent variable. Figure18 illustrates the experimental outcomes of wheat yield prediction based on rainfall.

Rainfall's complicated impacts on wheat output can vary depending on a number of variables, such as the quantity, timing, and dispersion of the precipitation. Soil pH influences several soil characteristics, all of which eventually impact yield production. Therefore, experiments were conducted using the collected data on soil pH in the Khairpur district. Figure19 illustrates theexperimental results of wheat yield prediction based on soil pH.

```
[2] import pandas as pd
    df = pd.read_csv('Humidity-Temperatur.csv', encoding = 'latin-1')
    df.head()
```

	Temperature ?C	Humidity	Crop Yield
0	44.39	87.6	362.53
1	36.17	87.9	368.21
2	32.51	87.5	378.16
3	37.56	86.9	361.98
4	45.71	87.4	385.02

Figure 10. Selected Input Values to Execute Operation of Correlation

```
[3] corr_matrix = df.corr()
    corr_matrix
```

	Temperature ?C	Humidity	Crop Yield
Temperature ?C	1.000000	-0.094129	0.157130
Humidity	-0.094129	1.000000	-0.295457
Crop Yield	0.157130	-0.295457	1.000000

Figure 11. Correlation between Humidity and Temperature Variables

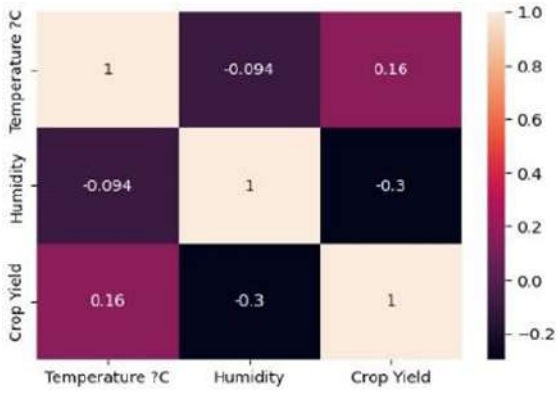


Figure 12. Explanation of Heatmap for Operation of Correlation

6.5 Multiple Linear Regression (MLR)

The multiple linear regression emerges using the least squares method. MLR is based on a number of fundamental assumptions, such as normality of errors, self-determination of errors, homogeneity

```
[ ] import pandas as pd
    df = pd.read_csv('Humidity-Temperatur.csv', encoding = 'latin-1')
    df.head()
```

	Temperature ?C	Humidity	Crop Yield
0	44.39	87.6	362.53
1	36.17	87.9	368.21
2	32.51	87.5	378.16
3	37.56	86.9	361.98
4	45.71	87.4	385.02

Figure 13. Selected Input Values for Correlation Between Rainfall and Humidity

(unchanged variability of irregularities), and linear behavior. Before conducting experiments with the linear regression model, preliminary data exploration operations are performed. The dataset, which includes six attributes (Year, Temperature, Rainfall, Humidity, pH Value, and Crop Yield), is depicted in Figure20. Therefore, in the model of MLR, the core objective

```
[3] import seaborn as sns
    import matplotlib.pyplot as plt
    corr_matrix = df.corr()
    corr_matrix
```

	Humidity	Rainfall (mm)	Crop Yield
Humidity	1.000000	1.000000	-0.295457
Rainfall (mm)	1.000000	1.000000	-0.295457
Crop Yield	-0.295457	-0.295457	1.000000

Figure 14. Correlation of Humidity and Rainfall with Crop Yield

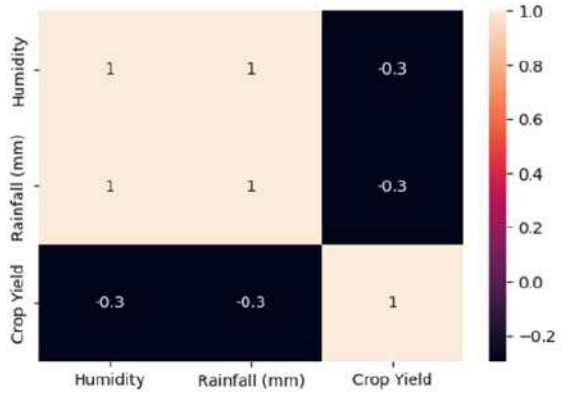


Figure 15. Explanation of Heatmap for Operation of Correlation

is to predict the 'Crop Yield' using the remaining characteristics.

The EDA technique is employed, as seen in Figure 21to determine whether the dataset contains any null values. The best course of action would be to remove that specific value if it is found. If null values account for 10% to 20% of the total dataset, methods such as using the mean, median, or mode are typically chosen to impute the missing values for the respective columns. However, when null values exceed 20%, the model risks losing crucial data, which can compromise the accuracy of predictions.

The quantitative data known as categorized data is composed of information that has been divided into variables that have been categorized into groups. Linear regression models assume that the input data is symmetrically distributed. Unlike basic linear regression models, various linear regression models employ various training parameters to identify the target parameter. Figure22 illustrates that in the multiple linear regression model, there are five coefficients corresponding to the five training features.

```
[6] plt.ylabel('Wheat Yield')
    plt.scatter(df.Temperature,df.Yield, color='red',marker='+')
```


Figure 16. Wheat Yield Prediction using Temperature

```
[5] plt.xlabel('Humidity')
    plt.ylabel('Wheat Yield')
    plt.scatter(df. Humidity, df.Yield, color='Red',marker='+')
```


Figure 17. Wheat Yield Prediction using Humidity

The framework trained to anticipate the targeted variable i.e crop productivity, as indicated in Figure23, using the previously specified training attributes, coefficient, and intercept parameters.

The coefficient of determination, denoted as R-squared (R2), is a statistical metric used in multiple linear regression. It measures the proportion of the variance in the dependent variable that can be explained by the independent variables included in the model. In multiple linear regression, the R2 value falls between 0 and 1, such asR2=0, it indicates that none of the variation in the response variable around its mean is explained by the model.

When R2=1, it signifies that all of the variance of the

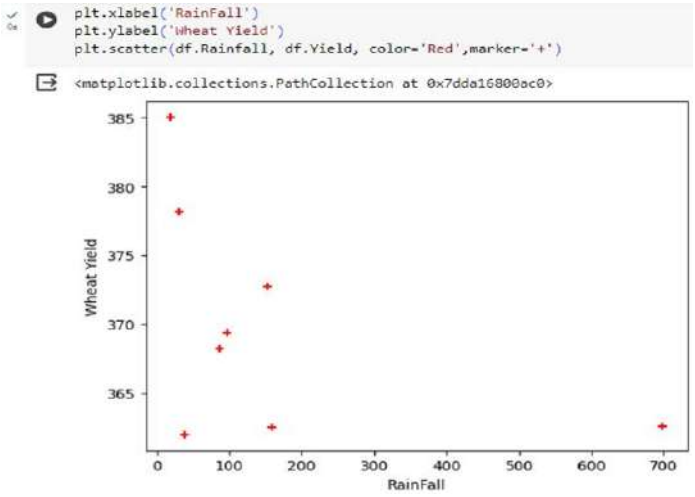


Figure 18. Wheat Yield prediction using Rainfall

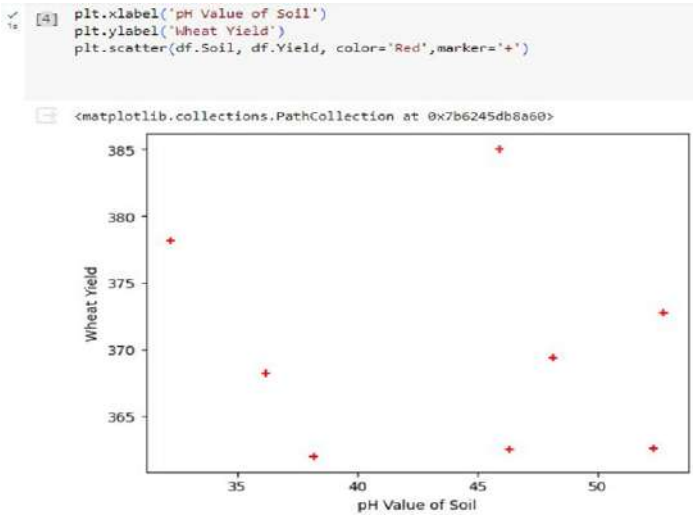


Figure 19. Wheat Yield Prediction using pH Value of Soil

response variable around its mean is explained by the model, meaning the model fits the data perfectly.

A high R2 value, which is in the neighborhood of 1, indicates that the factors that are independent in the model may account for a significant percentage of the volatility in the dependent variable.High H2 value highlights the model's performance at satisfactory level regarding identifying and exposing hidden patterns in the data. While the variation seen in the dependent variable against independent variables is highlighted by 1-score of R2 indicator. On the other hand, 0-score of R2 indicator highlights that the independent variables used in the regression model are failed at signifying dependent variable. Such results indicating the lacking of model against the dataset as well as use of insufficient and irrelevant independent variable. Hence, the calculating R2-score is essential in order to evaluate and validate the developed model in term of explaining variance in dependent variable. The resultant model's R2 value is "1.0," or 100% prediction accuracy, as shown in Figure24.

It is important to highlight that R2 alone does not indicate whether the model is correctly specified or if the coefficients are individually or cooperatively substantial. Therefore, to thoroughly assess the model's performance and validity, it's essential to employ additional diagnostic tools such as residual analysis and consider p-values alongside R2. These tools provide a more comprehensive evaluation of how well the model fits the data and whether the relationships between variables are statistically significant.

A residual in a case of linear regression is the variance among the dependent variable's recorded (real) value and the value anticipated by the regression model. The *i*-th observation's residual (e_i) can be computed mathematically as follows:

$$e_i = y_i - \hat{y}_i \tag{5}$$

Figure25 displays the residual that are computed for the developed model.

7 Results

The data for the independent variables, namely Rainfall, Temperature, pH Value of Soil and Humidity are collected individually. Initially, studies were conducted using SVR and SLR algorithms with different variables, which were referred to as five datasets.

7.1 Results with Dataset-I

For SVR and SLR models, datasets were separately created for each independent variable. Dataset-I specifically consists of temperature variable data. Low average temperature and high average temperature classifications have been generated from the temperature data in order to anticipate wheat crop production. When predicting yield in relation to elevated average temperature, Figure26 displays the calculated wheat yield prediction with average high temperature. The

	Year	Temerature	Rainfall	Humidity	pH Value	Crop Yield
0	2015	44.39	46.27	24.20	46.34	362.53
1	2016	36.17	35.81	26.03	36.20	368.21
2	2017	32.51	32.19	30.78	32.26	378.16
3	2018	37.56	38.06	26.90	38.21	361.98
4	2019	45.71	46.23	25.37	45.93	385.02
5	2020	48.23	48.51	24.29	48.12	369.41
6	2021	53.21	53.29	31.08	52.73	372.74
7	2022	52.32	52.29	43.50	52.33	362.56

Figure 20. Input Dataset based on Six Attributes

```
[19] i.isnull().sum()

Year           0
Temerature     0
Rainfall       0
Humidity       0
pH Value       0
Crop Yield     0
dtype: int64
```

Figure 21. Checking Null values using EDA Operation

```
[20] c = lr.intercept_

[21] c

     6222.943281622694

[22] m = lr.coef_
     m

     array([  -2.85461879,  -35.38359274,  229.49190271,    6.34561559,
             -200.90274011])
```

Figure 22. The Values of Coefficient and Intercept Used for Training

values of the intercept and coefficient are also used in experiments which are remain unchanged i.e., 5.11 and 158.44 respectively over the course of the studies, while the average high temperature value varies to determine different prediction outcomes.

In an identical manner, experiments with low mean temperatures are conducted and average temperature data for the previous seven years is also obtained. Figure27 presents the yield prediction results obtained using Low Average Temperature across

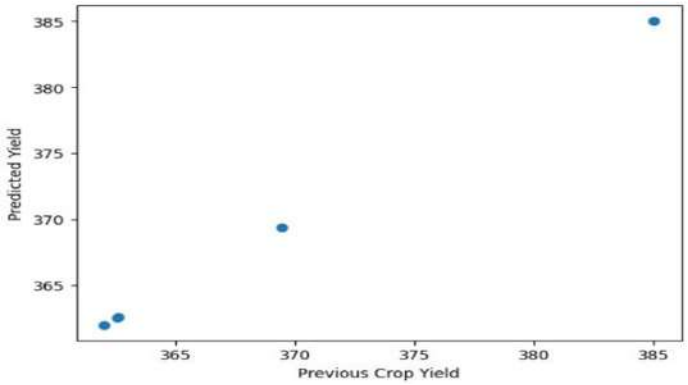


Figure 23. MLR Model Based Wheat Yield Prediction

```
[36] from sklearn.metrics import r2_score

[37] r2_score(y_test, y_pred_test)

     1.0
```

Figure 24. R2 Value of the Developed Model

the selected areas whereas the value of Coefficient is 3.43 and the value of Intercept is 276.88.

7.2 Results with Dataset-II

Many researchers have utilized simple linear regression for predicting crop yields using datasets from various crops. Similarly, in this study, wheat yield prediction is conducted using the rainfall data individually through a Simple Linear Regression model. The values of Coefficient and Intercept (-0.02 and 372.84) are used in the experiments with the dataset-II. The obtained findings of the rainfall and wheat crop yield forecast-

```
Residuals: 0    -1.368008
1     0.649269
2    -0.358771
3    -3.987938
4    11.878659
5    -3.927251
6    -4.355431
7     1.469470
Name: Crop Yield, dtype: float64
```

Figure 25. Residual Calculated for the Developed Model

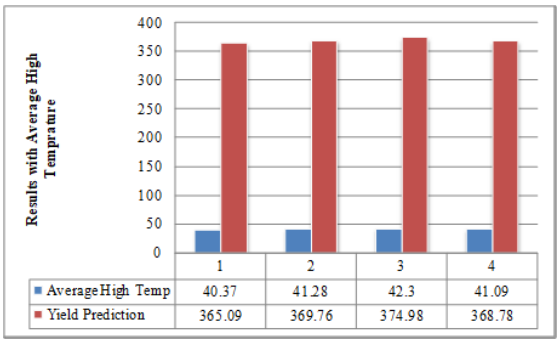


Figure 26. Measured Prediction Outcomes through Average High Temperature

ing for Taluka Gambat is depicted in Figure 28.

7.3 Results with Dataset-III

The third dataset only includes seven years' worth of humidity data. As a result, the model additionally trained to make the forecast using independent humidity data. The results obtained from this dataset are presented in Figure29. These outcomes are derived from the developed regression model that utilizes humidity as one of the main parameters for predicting yield. The collected and used values of Coefficient and Intercept are -0.28 and 378.27 respectively.

7.4 Results with Dataset-IV

Finally, a prediction was also made using the pH Value of Soil data from Dataset-IV. A comprehensive overview of the prediction results obtained with pH Value of Soil from Dataset-IV is shown in Figure 30. In the experimental purposes the values of Coefficient and Intercept are -0.15 and 377.09 respectively are used.

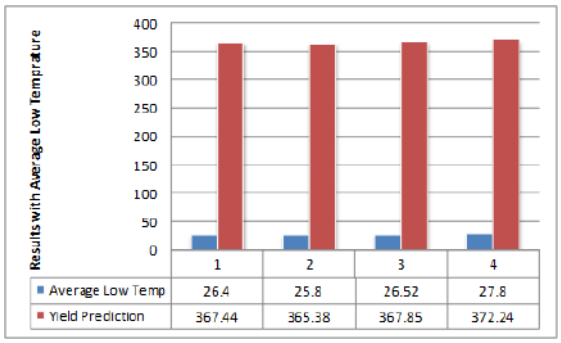


Figure 27. Measured Prediction Outcomes through Average Low Temperature

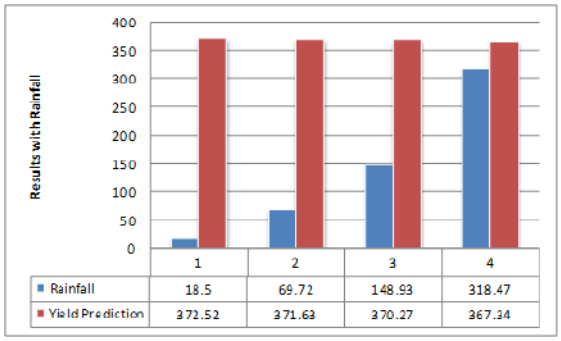


Figure 28. Measured Prediction Outcomes with Rainfall

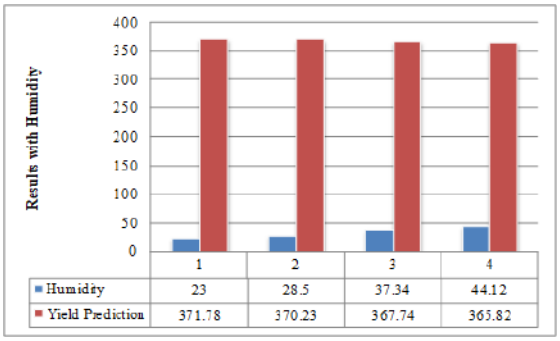


Figure 29. Measured Prediction Outcomes with Humidity

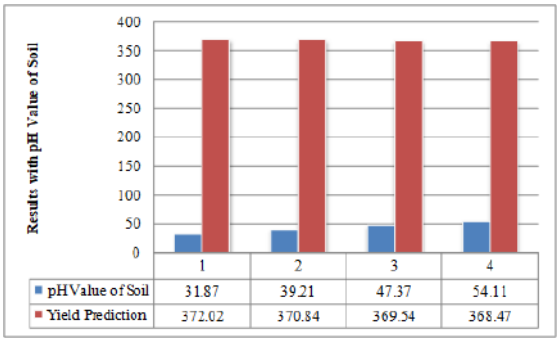


Figure 30. Measured Prediction Outcomes with pH Value of Soil

Table 3. Multiple Predictor Variables Based Yield Prediction Results

Predictor Variable	Coefficient value	Intercept	R-square	Residual
Temp	3.59	285.44	1	4.73
Humidity	2.98	273.92	0.92	1.18
Rainfall	3.22	264.22	0.79	1.57
pH-value of Soil	3.43	279.21	1	3.91

7.5 Results with Dataset-V

The SVR and MLR regression models are trained using multiple parameters to achieve optimal prediction outcomes. These models utilize all independent variables — Temperature, Rainfall, Humidity, and pH value of Soil — collectively during training. When predicting with single predictor variables, suitable outcomes were observed with Dataset-I and Dataset-IV. However, the results obtained with Dataset-II and Dataset-III were not satisfactory. For this purpose, in order to make prediction using all of the selected variables combined, multiple linear regression models and SVR with multiple predictor variables were also constructed, producing very desirable results. Table 3 discusses the obtained results using various models.

8 Conclusion

In today's context, predicting crop yield is crucial amidst global climate change. Initially, variables chosen for regression analysis experiments include Temperature, Rainfall, Humidity, and Soil pH, with wheat yield production per year selected as the dependent variable. Data collection focused on District Khairpur, specifically Taluka Gambat, gathered from various sources such as official departments, news articles, research papers, and websites. To develop the wheat yield prediction model, support vector, simple, and multiple linear regression approaches are applied. Collinear associations across the independent and dependent variables are discovered for the purpose of model validation. Five datasets were prepared for each selected variable, facilitating experiments aimed at enhancing accuracy in predicting wheat yield. Each dataset was split into 70% for training and 30% for testing the developed prediction model. Statistical results were then evaluated using multiple linear regression models.

In this paper, the data is collected from Taluka Gambat. In future, the data of other Talukas of District Khairpur will be gathered and experimented. Only Wheat crop is preferred in this research. The Dates, Guava, Rice and some other crops and fruits will be considered for yield prediction and data will be statistically analyzed. The statistical technique of regression analysis is used in this paper, in future some advanced techniques will be used as given in References [23] [24] for getting more authentic and reliable prediction results.

Author Contributions

Abdul Sami Kharal: Conceptualization and designed the study, developed the methodology, conducted the experiments, and analyzed the data **Shahid Ali Mahar:** assisted in refining the methodology, provided critical feedback on experimental design, and contributed to data analysis. **Muhammad Imran Mushtaque:** Visualization, Exploring of data **Aurangzeb Magsi:** Software, Validation. **Javed ahmed Mahar:** contributed to the interpretation of results and the writing of the manuscript.

Compliance with Ethical Standards

The authors declare no conflict of interest. This article does not involve studies with human participants or animals conducted by any of the authors. Informed consent was obtained from all individual participants included in the study.

References

[1] Maria Sassi. Economic connectiveness and pro-poor growth in sub-saharan africa: The role of agriculture. *Sustainability*, 15(3):2026, 2023.

[2] Bushra Zaman, Muhammad Niamat Ullah, Saima Munir, and Shehla Mazhar. Impact of agriculture, manufacturing, commodity producing sectors and service sectors on gdp growth rate: Empirical evidence from pakistan 1994-95 to 2021-22. *Al-Qanṭara*, 9(4):764–779, 2023.

[3] Muhammad Islam, Syed Ijaz Hussain Shah, Syeda Amna Wajahat, Muhammad Faheem Bhatti, and Noor Ul Ain. Analyzing the impact of subsectors and population growth on agricultural sector in pakistan. *STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH*, 5(2):29–37, 2023.

[4] Muhammad Shakeel, Najam ul Hassan, Kalsoom Akhtar Chaudhry, Muhammad Naveed Tahir, et al. What affects crop production in pakistan: The role of agriculture employment, machinery and fertilizer consumption. *Bulletin of Business and Economics (BBE)*, 12(3):541–546, 2023.

[5] Jaskaran Singh Madray. Crop cutting experiments of rabi oilseeds & wheat crops in punjab: A comparative study. *Journal of Composition Theory*, 13(3):416–421, 2020.

[6] BM Nayana, Kolla Rohit Kumar, and Christophe Chesneau. Wheat yield prediction in india using principal component analysis-multivariate adaptive regression splines (pca-mars). *AgriEngineering*, 4(2):461–474, 2022.

[7] Daniela De Benedetto, Emanuele Barca, Mirko Castellini, Stefano Popolizio, Giovanni Lacolla, and Anna Maria Stellacci. Prediction of soil organic carbon at field scale by regression kriging and multivariate adaptive regression splines using geophysical covariates. *Land*, 11(3):381, 2022.

[8] EMP Ekanayake, LCD Wickramasinghe, and RT Weliwatta. Use of regression techniques for rice yield estimation in the north-western province of sri lanka. *Ceylon Journal of Science*, 50(4):439–447, 2021.

[9] Surabhi Lingwal, Komal Kumar Bhatia, and Manjeet Singh. A novel machine learning approach for rice yield estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(3):337–356, 2024.

[10] Yulin Shen, Benoît Mercatoris, Zhen Cao, Paul Kwan, Leifeng Guo, Hongxun Yao, and Qian Cheng. Improving wheat yield prediction accuracy using lstm-rf framework based on uav thermal infrared and multispectral imagery. *Agriculture*, 12(6):892, 2022.

[11] Purbasha Mistry and Ganesh C Bora. Development of yield forecast model using multiple regression analysis and impact of climatic parameters on spring wheat. *International Journal of Agricultural and Biological Engineering*, 12(4):110–115, 2019.

[12] Shurong Yang, Lei Li, Shuaipeng Fei, Mengjiao Yang, Zhiqiang Tao, Yaxiong Meng, and Yonggui Xiao. Wheat yield prediction using machine learning method based on uav remote sensing data. *Drones*, 8(7):284, 2024.

[13] M Jawed Iqbal, Zaeem Uddin Ali, and S Shahid Ali. Agroclimatic modelling for estimation of wheat production in the punjab province, pakistan. *Proc. Pak. Acad. Sci*, 49: 241–249, 2012.

[14] K Akhand, M Nizamuddin, and L Roytman. *Global Journal of Science Frontier Research: D Agriculture and Veterinary*, 18(2):1–8, 2018.

[15] IA Suprio, JA Mahar, and A Maitlo. Regression analysis of rice data for yield prediction using python programming language. *Sindh University Research Journal (Science Series)*, 51(2):183–188, 2019.

[16] IA Suprio, JA Mahar, and SA Mahar. Rice yield prediction and optimization using association rules and neural network methods to enhance agribusiness. *Indian Journal of Science and Technology*, 13(13):1367–1379, 2020.

[17] F Aslam, A Salman, and Jan. I. Predicting wheat production in pakistan by using an artificial neural network approach. *Sarhad Journal of Agriculture*, 35(4):1054–1062, 2019.

[18] L.O. Joel, W. Doorsamy, and B.S. Paul. A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 5(3):971–1005, 2022.

[19] S. Karimzadeh and S. Olafsson. Data clustering using proximity matrices with missing values. *Expert Systems with Applications*, 126:265–276, 2019.

[20] K Murakami, S Shimodal, Y Kominami, M Nemoto, and S Inoue. Prediction of municipality-level winter wheat yield based on meteorological data using machine learning in hokkaido, japan. *PLoS ONE*, 16(10):24–32, 2021.

[21] I. Kamenova, M. Chanev, P. Dimitrov, L. Filchev, B. Bonchev, L. Zhu, and Q. Dong. Crop type mapping and winter wheat yield prediction utilizing sentinel-2: A case study from upper thracian lowland, bulgaria. *Remote Sensing*, 16(7):1144, 2024.

[22] T.H. Kim, V.S. Solanki, H.J. Baraiya, A. Mitra, H. Shah, and S. Roy. A smart, sensible agriculture system using the exponential moving average model. *Symmetry*, 12(3): 457, 2020.

[23] A Ahmed, I Parveen, S Abdullah, I Ahmad, N Alturki, and L Jamel. Optimized data fusion with scheduled rest periods for enhanced smart agriculture via blockchain integration. *IEEE Access*, 12:15171–15193, 2024.

[24] A Ahmed, S Abdullah, M Bukhsh, I Ahmad, and Z Mushtaq. An energy-efficient data aggregation mechanism for iot secured by blockchain. *IEEE Access*, 10:11404–11419, 2022.

[25] K. Sahoo, A.K. Samal, J. Pramanik, and S.K. Pani. Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, 8 (12):4727–4735, 2019.