International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Prediction of Soil Macronutrients Using Machine Learning Algorithm

Umm E Farwa[a*], Ahsan Ur Rehman[b], Saad Qasim Khan[c], Muhammad Khurram[d]

[a,b]*Smart City Lab – NCAI, Computer & Information Systems Engineering Department, University Road Gulshan-e-Iqbal, Karachi 74600, Pakistan*

[c,d]*Computer & Information Systems Engineering Department, University Road, Gulshan-e-Iqbal, Karachi 74600, Pakistan*

[a]*Email: umm-e-farwa@hotmail.com*

[b]*Email: ahsan191@hotmail.com*

[c]*Email: saadqasimkhan@neduet.edu.pk*

[d]*Email: mkhurrum@neduet.edu.pk*

Abstract

In this research work, machine learning algorithms were applied to find the relationship between independent variables and dependent variables for soil data analysis. The independent variables include moisture, temperature, soil pH, Cation Exchange Capacity(CEC) whereas, the dependent variables include Nitrogen, Phosphorus and Potassium (NPK). This research concludes relationships between Phosphorus, Potassium, soil pH and CEC; Nitrogen and soil moisture and temperature using machine learning(ML) algorithms so as to deduce NPK content of soil. A comparative analysis with obtained results from each ML method is also presented. Machine learning algorithms are best performed on data with multiple independent variables. The values computed for nitrogen relationship were more accurate than PK relationship values. The accuracy of data set I was less than data set II. A large data set would produce more accurate results for both data sets.

Keywords: Machine Learning(ML); Cation Exchange Capacity(CEC); Linear Regression; Ridge Regression; Bayesian Regression; NPK; Agriculture; Soil Nutrients; Soil pH; Nitrogen; Phosphorus; Potassium.

---

* Corresponding author.

1. Introduction

Soil nutrients are an essential part of agriculture. Farmers study several aspects of the soil before farming. Soil Nitrogen, Phosphorus and Potassium (NPK) is an important parameter that helps in the growth of the plant [1]. It is important to understand the NPK values before implication of fertilizers over the field. Nitrogen helps in the growth of leaves on the plant, Phosphorus contributes to root growth, flower growth and fruit development while Potassium helps overall functions of the plant [2]. It is important to understand the relationship between soil pH and NPK. If an equation is present, then the farmer will accurately use the fertilizer. The crop yield will increase, and the efficiency will improve. Machine learning is used to identify patterns. It derives the relationship between independent variables and dependent variables. Nitrogen, Phosphorus and Potassium (NPK) values are dependent variables [3]. They are dependent on soil pH, moisture, temperature and cation exchange capacity. Machine learning methods were applied on soil nutrient samples for pattern recognition in soil datasets. Researchers discovered nitrogen fertilizer in the early 20^{th} century. However, basic understanding of soil nutrients even before the 20^{th} century was lacking. The farmers never took complete advantage of this information. The last three decades have changed the agriculture world. Farmers in developed countries have extensively used lab testing to find NPK values [4]. The NPK ratio of the soil informs the farmers about the ratio of the fertilizer. For example, if the soil has high amounts of Phosphorus and Potassium, but not nitrogen, then farmers shall use a fertilizer that has more nitrogen and less Phosphorus and Potassium. However, farmers in emerging countries still have no idea how to use the information to their advantage. The fertilizer companies print NPK values on fertilizer bags. However, farmers are unsure of how to use the values correctly. It is impossible to find NPK values with the naked eye. Detailed analysis is required to understand the NPK values. Therefore, farmers often misuse fertilizers, and as a result, the crop yield is reduced, or it is not perfect. Every plant has a different NPK requirement [5]. Some plants require more nitrogen, while others require more potassium and phosphorus. The lack of data on soil nutrients forces farmers to plant crops on unsuitable soil, and they eventually misuse the fertilizer as they have no idea about the soil nutrients. The farmers mismanage land and other precious resources such as. As a result, developing countries like Pakistan suffer from water shortage and low crop yield. The misuse of resources puts a strain on food security in the region. Pakistan is an emerging country. It has a large agriculture sector. However, the industry is not even close to its true potential [6]. Pakistan produces large quantities of different crops. The country still has low exports and high imports. The main reason is the low- quality crops. It is mind-boggling that a country with 47% of agricultural land suffers from low quality of crops [7]. The farmers in Pakistan do not know how to calculate NPK values. They are not aware of how NPK values affect crops. Therefore, they inappropriately use the fertilizer and as a result, crops suffer [8]. Pakistan has the potential to lift its economy through agriculture. If the crop yield is anywhere near the maximum potential, then the country will benefit from the high exports. Also, the growth in agricultural production would provide sufficient raw materials to the industrial sector of the country.

1.1. Related work

Researchers have previously used the Bayesian method to find the relationship between independent variables and dependent variables. Previous researchers used various machine learning methods to create a comparative analysis. The previous analysis will help the research to progress further. Researchers performed a detailed

analysis of the soil nutrients to find the relationship between all variables [9]. However, it was a complex task, and the samples were not sufficient enough to find a reliable value. This research will focus on Nitrogen, Phosphorus and Potassium values and their relationship with pH and CEC. It is easier to understand one relationship before trying to study in-depth soil data. Lab-testing was the traditional method before the arrival of advanced technology. The lab would send the test results back to the farmers. The farmers used the correct ratio of NPK in fertilizer due to the lab reports. Currently, several advanced devices take values directly from the soil [10]. These devices need to physically sample the soil and process the values. Other researchers have used machine learning to predict the values of soil nutrients. Previous research focused on the development of hardware that contains sensors to take soil pH values [11]. The hardware is necessary to acquire the readings. However, it is essential to find the relationship between soil nutrients and soil pH. The Cation Exchange Capacity (CEC) is related to nutrients values as well as pH. Therefore, it is possible to find the relationship between these variables. This research utilizes the samples taken from another research and aims to find a relationship between soil acidity, conductivity and soil nutrients. Different regression techniques were used to derive regression models. The expected value was found to be a combination of linear features. Therefore, the regression models were ideal for the development of the solution. The research aims to propose a comparative analysis to identify the most effective and accurate method. The machine learning models were approximated over a dataset of more than 900 soil samples. In the study, we have used the following regression methods to find how pH and CEC affect phosphorus and potassium contents in soil and how soil moisture- temperature affects soil nitrogen. We used several methods that perform L1 and L2 regularisation. L1 regularisation is a mathematical function that decreases the absolute difference sum [12]. It decreases the sum of absolute differences between the estimated value and the target value. L2 regularisation is similar to L1 regularization. However, it minimizes the sum of the square of differences.

1.2. Solution

Machine learning provides an accurate relationship between soil nutrients. The researchers will help the farmers to plant crops that are suitable for the soil. The crop yield will significantly increase. The efficiency of the whole process will increase. Machine learning will help farmers to use fewer resources to produce more crops [13]. It is a great opportunity for farmers to lower the expenditure while increasing the crop yield. We will obtain the solution through the use of machine learning methods such as ridge regression, linear regression and Bayesian regression. The research will only present the relationship of PK with CEC and pH as the researchers were unable to gather nitrogen values in the same data set.

1.3. The objective of the research

The objective of this research is to find the relationship between soil pH-CEC and PK. The knowledge about the soil will empower the farmers to better use the fertilizer [14]. The understanding of the relationship between soil pH-CEC and PK will enable the farmers to increase the crop yield. The research aims to find the relationship between NH4, NO3 and moisture-temperature. The first dataset had no nitrogen values. Therefore, the research will find the relationship between nitrogen and soil moisture- temperature separately. Machine learning will provide an equation that defines the relationship between independent variables and dependent variable. The

equation will eliminate the need for physical testing of NPK values. The farmers will require soil pH and CEC values. Machine learning methods will predict PK values. It will enable farmers to use the correct fertilizer without extensive testing and expenditure.

2. Implementation

2.1. Soil dataset for analysis

Two datasets containing soil attributes were collected online from Kellogg Biological Station-Long Term Ecological Research (KBS LTER). Dataset 1 comprises of the following attributes, pH, CEC, Phosphorus, Potassium, Magnesium, Calcium and Lime Index [15]. Dataset 2 comprises of the following attributes: Soil Nitrate (NO3), Soil Ammonium (NH4+), Soil Moisture and Soil Temperature [16].

2.2. Linear regression

Linear regression helps in understanding the relationship between one dependent variable, and one or more independent variables. The researchers used the linear regression model to find an equation that accurately defines the relationship between pH, CEC and different elements of NPK. The use of multiple independent variables is known as multiple linear regression [17]. In the study, we have used both linear regression and multiple linear regression to find how pH and CEC affect nitrogen, phosphorus and potassium. The machine learning model was given approximately 900 soil samples. The equation for simple linear regression is given below (1):

Equation 1

$$\hat{y}\,(w, x) = w0 + w1x1 + \ldots + wp \times p \qquad (1)$$

$\hat{y}$ is the dependent variable while x is the independent variable. Across the module, we designate the vector w = (w1,..., wp) as coefficient vector and w0 as function intercept.

2.3. Ridge regression

We used ridge regression to analyse the data. The use of ridge regression was important to eliminate multicollinearity [18]. It is essential to use ridge regression when the linear relationship is near perfect. The regression causes division by zero if the linear relationship is near perfect. The division by zero aborts all calculations. As a result, we are unable to calculate the relationship. Therefore, we used ridge regression to better define the relationship between independent variables and dependent variables. The ridge regression makes coefficients strong against collinearity. Equation. (2) shows the relation.

Equation 2

$$L_{ridge}\,(\hat{\beta}) = \sum_{i-1}^{n} (y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m} \hat{\beta}_j^2$$

3. Best Machine Learning Algorithm

3.1. Bayesian regression

The Bayesian regression provides a probabilistic model of a regression problem. It includes the regularization parameters in the estimation procedure [19]. The Bayesian regression is flexible as it can adapt to data. The regularisation parameters help in estimating a probabilistic model. We have used Bayesian regression to estimate future values. It is necessary to create a probabilistic model to estimate future values. We used Bayesian regression to create a probabilistic model. The other regression methods are useful in finding relationship, but the Bayesian regression allows the researchers to create a probabilistic model that can predict relationships in future. Bayesian learning is more powerful than other machine learning models. The probabilistic model incorporates previous results with current data. Therefore, it creates a more accurate model. We choose this method to prevent overfitting. Overfitting occurs when there is a small dataset. The machine learning method only learns the relationship between specific variables, but it is unable to adjust to unseen values. However, in Bayesian regression, prior knowledge helps to create a better model. Bayesian regression is more complex than linear regression. The method does not find a single estimated value. It creates a probabilistic model to resolve unexpected values [20]. We want to create an equation that represents the relationship between pH- CEC and NPK. Bayesian regression is resolving all the unknown values. It is reducing the uncertainty within the model. For example, a family dines thrice a week in a restaurant, but they do not visit the restaurant on any specific days. The visits are random, and they come as they please. Bayesian regression will use the data of previous visits to determine the days when the family is more likely to dine in the restaurant. The values are not completely accurate, but it reduces randomness from a data set. Therefore, we used Bayesian regression to remove the randomness from our model. Equation 3. It shows the influence of priors and likelihood and its accumulative effect on proceeding posterior values with normalization.

$$Likelih\,ood \times Prior \qquad (3)$$

$$Posterior = \frac{}{Normalization}$$

Posterior refers to the statistical probability of a hypothesis. The hypothesis is based on statistical observations. Therefore, posterior predicts new values for future events based on previous observations. The likelihood is different from posterior; it refers to an event in the past. It expects that the same event in the past would produce the same values in future. Prior refers to model parameters that the programmer enters. These parameters are the expected values. Unlike other models, Bayesian regression allows researchers to add values that they expect. Normalization adjusts differently measured values. It is important to adjust values to improve the alignment of values. New observations or data set can improve the existing model. Therefore, we choose this model as we can incorporate new data in the model. Ideally, we will input future data sets in the model to improve the model.

3.2. Bayesian ridge regression

Bayesian Ridge follows the same methodology as Bayesian so as to commute a probabilistic model for a

regression problem. The priors over α and λ are selected to be gamma distributions, the conjugate prior for the precision of the Gaussian. The resulting model is called Bayesian Ridge Regression, and is similar to the classical Ridge. The parameters w, α and λ are estimated jointly during the fit of the model, the regularization parameters α and λ being estimated by maximizing the log marginal likelihood.

The remaining hyperparameters are the parameters α1, α2, λ1 and λ2 of the gamma priors over α and λ. These are usually chosen to be non- informative. The default value of these hyperparameters are as follows α1=α2=λ1=λ2=10−6.

3.3. Machine learning model

A learning model may consist of several modules based on the complexity of the data under consideration. A typical learning model is divided into three main modules as shown in Fig. 1 such as training data repository which holds the dataset for training the model and the related metadata information, the second module is the learning unit that is programmed with machine learning algorithms and statistical principles so as to perform data analysis and computes a learning model. The third module is data prediction or pattern recognition module which is responsible for generating results on the basis of derived data models. Initially, a training dataset is chosen after the application of data normalization techniques such as outlier detection, data screening and filtration so as to eliminate the erroneous data points within the dataset. This normalized data is then sent to the learning unit where different algorithms are applied on the training dataset so as to derive data patterns and finally generate model equations for each corresponding learning algorithm. These algorithms are the different regression and data regularization techniques proposed throughout the world for solving data prediction complications.

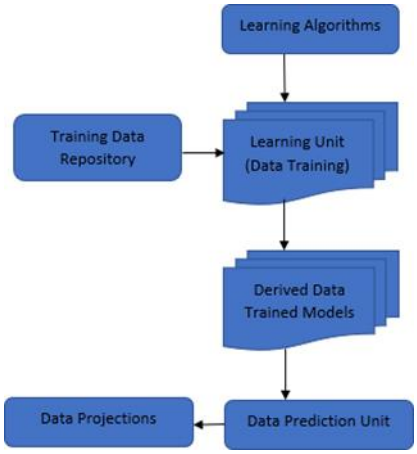

Figure 1: Machine Learning System Architecture

4. Observation

The research presents a comparative analysis of different regression methods. Therefore, the research will show the best results that were obtained. It is essential to lessen the error to find an accurate relationship. Since NPK values are widespread over the respective datasets. Therefore, we implemented the three regression methods to

identify relationship between soil parameters- pH, CEC and P K values while the influence of Soil Moisture and Temperature on Soil Nitrogen- NO3 and NH4 was also concluded. Machine learning offers various solutions to solve a prediction problem. We used linear multiple, ridge and Bayesian techniques and found that ridge regression models performed better among the three techniques with the training size of 0.9 and test size of 0.1 of the datasets. The figures illustrate pattern recognition of NPK values from the two soil datasets when ridge regression was applied. Figure.2 and Figure. 3 illustrates that the model predicted K value quite accurately as compared to the predicted value of P which is also evident from R2 values. Figure.4 and Figure. 5 illustrates that the model predicted NO3 value quite accurately as compared to the predicted value of NH4 which is also evident from R2 values.

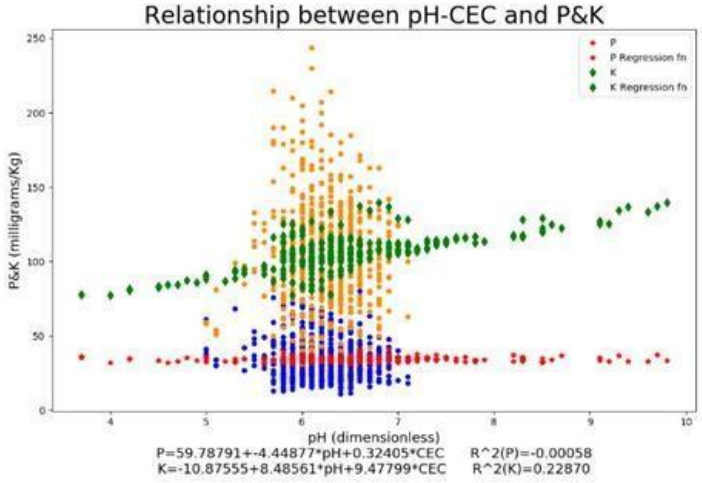

Figure 2: Ridge regression model of pH with corresponding CEC vs P(Phosphorous) and Potassium(K)

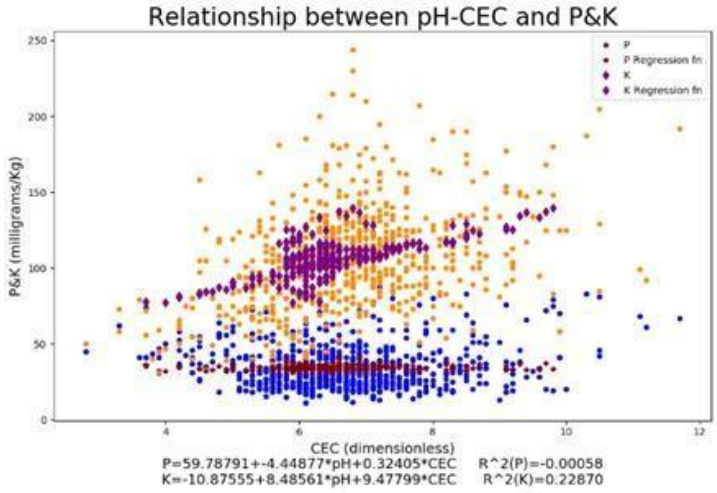

Figure 3: Ridge regression model of CEC with corresponding pH vs P(Phosphorous) and Potassium(K)

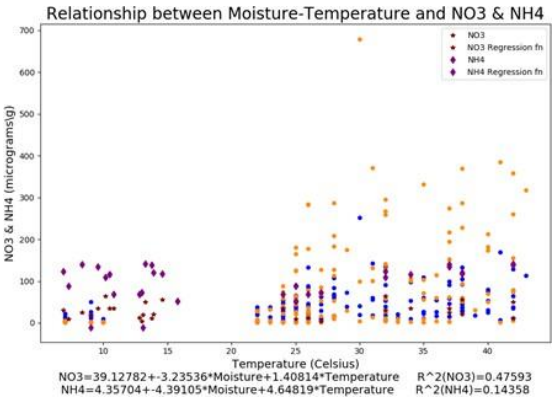

Figure 4: Ridge regression model of Temperature with corresponding Soil Moisture vs NO3 (Soil Nitrate) and NH4 (Soil Ammonium)

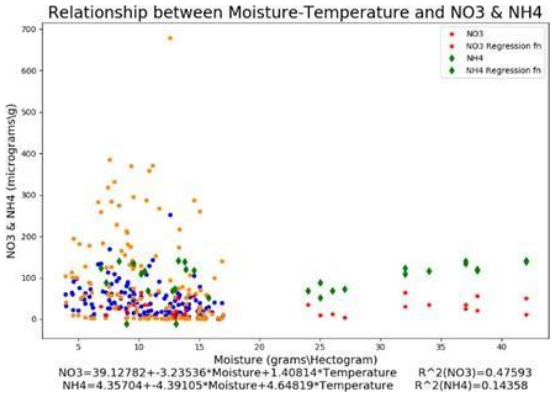

Figure 5: Ridge regression model of Soil Moisture and Temperature vs NO3 (Soil Nitrate) and NH4 (Soil Ammonium)

5. R Square Results

The results show that the output is better when there are multiple inputs. Therefore, we obtained better equations with the use of multiple inputs. It is better to find the relationship between pH-CEC and Potassium or Phosphorus. The machine learning algorithms worked better when there were two independent variables and one dependent variable. R square represents how close data is to the regression line. The R square refers to the coefficient of determination [21]. It shows the relationship between the variability of data. Therefore, a 100% score means that two data points are correlated to each other. In contrast, a 0% score means no relationship affects variability.

Table 1: R Square Regression Values for dataset I

Linear Regression	
For P, K using pH:	
P = 47.96816 + (– 1.78966) pH	R^2 (P)= - 0.03675
K= 30.42523 + (12.49256) pH	R^2 (K) = 0.00658

For P, K using CEC:	
P = 30.32063 + (0.54455) CEC	R^2 (P) = 0.00225
K = 37.36155 + (10.08913) CEC	R^2 (K) = 0.11554
For P, K using pH & CEC:	
P = 55.57924 + (-4.00611) pH + (0.63728)CEC	R^2 (P)= -0.11630
K = (-12.44206) + (7.84011) pH + (10.19458)CEC	R^2 (K)= 0.11203
Ridge Regression	
For P, K using pH:	
P = 48.81444 + (– 2.09660) pH	R^2 (P)= - 0.02165
K= 32.74400 + (12.22071) pH	R^2 (P)= - 0.02165
For P, K using CEC:	
P = 30.15110 + (0.63068) CEC	R^2 (P) = -0.00634
K = 36.82853 + (10.19354) CEC	R^2 (K) = 0.05355
For P, K using pH & CEC:	
P = 59.78791 + (-4.44877) pH + (0.32405)CEC	R^2 (P)= -0.00058
K = (-10.87555) + (8.48561) pH + (9.47799) CEC	R^2 (K)= 0.22870
Lasso Regression	
For P, K using pH:	
P = 36.17962 + (– 0.0) pH	R^2 (P)= - 0.00052
K= 83.70622 + (3.98266) pH	R^2 (K) = 0.00791
For P, K using CEC:	
P = 34.05672 + (0.0) CEC	R^2 (P) = -0.00064
K = 48.92058 + (8.46692) CEC	R^2 (K) = 0.18831
For P, K using pH and CEC:	
P = 33.93717 + (-0.0) pH + (0.0) CEC	R^2 (P)= -0.00405
K = (45.22681) + (0.0) pH + (8.89141) CEC	R^2 (K)= 0.15007
Elastic Net Regression	
For P, K using pH:	
P = 36.46046 + (– 0.0) pH	R^2 (P)= - 0.00335
K= 98.89240 + (1.47541) pH	R^2 (K) = -0.00183
For P, K using CEC:	
P = 30.08235 + (0.58506) CEC	R^2 (P) = -0.01891
K = 56.26985 + (7.17895) CEC	R^2 (K) = 0.08161
For P, K using pH and CEC:	
P = 34.59389 + (-0.07448) pH + (0.0) CEC	R^2 (P)= 0.00024
K = (58.29564) + (0.09422) pH + (7.08763) CEC	R^2 (K)= 0.10280
Bayesian Ridge Regression	
For P, K using pH:	
P = 39.10180+ (– 0.50775) pH	R^2 (P)= - 0.00769
K= 38.19544 + (11.35910) pH	R^2 (K) = 0.00572
For P, K using CEC:	
P = 33.88129 + (0.06613) CEC	R^2 (P) = -0.00419
K = 47.02549 + (8.73303) CEC	R^2 (K) = 0.19877
For P, K using pH and CEC:	
P = 41.13913 + (-2.04105) pH + (0.86189) CEC	R^2 (P)= -0.02234
K = (1.68341) + (5.67431) pH + (10.13588) CEC	R^2 (K)= 0.06137

Table 2: Results from Dataset II for all 5 techniques

Linear Regression	
For NO3, NH4 using Soil Moisture and Temperature:	
Soil NO_3 = 46.38793 +(- 3.63243 M) + 1.26699 T	R^2 (Soil NO_3) = 0.05743
Soil NH_4 = 12.66578 + (- 4.50947 M) + 4.36830 T	R^2 (Soil NH_4) = 0.46669
Ridge Regression	
For NO3, NH4 using Soil Moisture and Temperature:	
Soil NO_3 = 39.12782 + (- 3.23536 M) + 1.40814 T	R^2 (Soil NO_3) =0.47593
Soil NH_4 = 4.35704 + (- 4.39105 M) + 4.64819 T	R^2 (Soil NH_4) = 0.14358
Lasso Regression	
For NO3, NH4 using Soil Moisture and Temperature:	
Soil NO_3 = 35.46984 + (2.38278 M) + 1.22019 T	R^2 (Soil NO_3) = 0.30432
Soil NH_4 = (- 22.53155) + (-1.18580 M) + 4.31183 T	R^2 (Soil NH_4) = 0.17622
Elastic Net Regression	
For NO3, NH4 using Soil Moisture and Temperature:	
Soil NO_3 = 45.02916 + (- 3.36862 M) + 1.26625 T	R^2 (Soil NO_3) = 0.25870
Soil NH_4 = 14.61092 + (- 5.17290 M) + 4.67543 T	R^2 (Soil NH_4) =-0.25617
Bayesian Ridge Regression	
For NO3, NH4 using Soil Moisture and Temperature:	
Soil NO_3 = 38.59602 + (- 3.07339 M) + 1.41238 T	R^2 (Soil NO_3) =0.11390
Soil NH_4 = (- 1.47765) + (- 4.48592 M) + 4.84721 T	R^2 (Soil NH_4) =0.11957

The table shows that single independent variable relationships were less accurate than compound relationships. Bayesian regression and ridge regression were the most accurate methods. These methods produced a better variability percentage. The research was limited by the number of soil samples. The accuracy of the methods will increase with more samples. Nine hundred samples are not enough for a machine learning algorithm. However, the algorithms still produced a variability percentage. R square was not 0% in any case. Therefore, these variables are related to each other. They are not perfectly correlated, but they are related.

Table 2: R Square and Square Root Values

	R^2(P)	R^2(K)	R^2(NO3)	R^2(NH4)
Linear Regression	-0.11630	0.11203	0.05743	0.46669
Ridge Regression	-0.00058	0.22870	0.47593	0.14358
Bayesian Regression	-0.02234	0.06137	0.11390	0.11957
	R(P)	|R(K)|	R(NO3)	R(NH4)
Linear Regression	0.34102	0.33470	0.23964	0.68314
Ridge Regression	0.0240	0.47822	0.68987	0.37891
Bayesian Regression	0.14946	0.24772	0.33749	0.34578

The regression methods were not highly accurate. The main reason for the lack of accuracy is the small number of samples. The data set I only had 940+ samples. Some of the samples were null values. Therefore, regression methods did not perform well. The regression method defined a relationship between variables. These variables

are related. Therefore, the accuracy of the system will improve with more samples.

Table 3: Most accurate R square results of NO3 and NH4 are highlighted in green

X		Linear Regression	Ridge Regression	Lasso Regression	Elastic-Net	Bayesian Ridge
Moisture & Temperature	**R^2(NO3)**	0.05743	0.47593	0.30432	0.25870	0.11390
Moisture & Temperature	**R^2(NH4)**	0.46669	0.14358	0.17622	-0.25617	0.11957

The data set II was used to find the relationship between NO3, NH4 and Moisture-Temperature. The first data set had no nitrogen values. Therefore, it was necessary to obtain nitrogen values from its compound form. It is more difficult to find the percentage of nitrogen, but the machine learning method was more accurate in finding the relationship between NO3, NH4 and Moisture-Temperature.

6. Limitations of the Study

Data was a huge limitation during the research. It is quite difficult to collect data in Pakistan due to several problems. However, the team collected data in Pakistan, and the team created a small dataset. The dataset was small because the team took a handful of samples. As a result, the sample size was too small for the algorithms. Machine learning algorithms require huge datasets to function for best results. So, the algorithms shall perform better when feeded with large datasets . The lack of human resources was an issue in the research. Researchers had no human resources to collect data. Therefore, they gathered a small number of samples as they did not have enough time. Furthermore, researchers conducted the research themselves without the help of interns or any other resource. Consequently, they performed menial and time-intensive tasks such as data collection. Data collection is a task for entry-level employees or interns. Since the researchers performed data collection, they

had less time to conduct research. As a result, each researcher was exhausted. So, the lack of human resources severely affected the research. Besides human resources and data collection, the data outliers were a problem for the research. Data outliers refer to incorrect data samples. Data collectors sometimes collect the wrong data, or they taint the data. As a result, the sample does not correctly represent the properties of the soil. Overall, it affects the value of the coefficient of determination. Since the team used data available on the internet, they had no method to remove outliers. Furthermore, their own data set was too small, and it had several outliers. So, the data outliers limited the quality of the research. The team had no reliable NPK sensors to collect data. Furthermore, most of the data samples were inaccurate. Consequently, the numbers of local data samples were low. Therefore, the lack of accurate equipment prevented researchers from collecting local data. As a result, the researchers did not invest time to collect data since they knew the sensors were inaccurate. So, unreliable equipment wasted the time of researchers, and it provided a few accurate samples. Consequently, it limited the research team to datasets on the internet.

7. Recommendations for the Future

The machine learning algorithms worked in this research. They showed the relationship between independent variables and the dependent variable. However, the relativity values were low since the number of data samples was low. In future, researchers should focus on data collection. The machine learning algorithm will work a lot better with bigger local datasets. Also, the local datasets are more accurate than the datasets available on the internet. So, researchers should find a way to collect big data in the future. The data will significantly improve the findings of this research. Furthermore, agriculture data will help the country to improve its farms. Data collection is an issue as NPK soil sensors are quite expensive. Therefore, researchers should find a cheaper alternative to collect data. However, the alternative should provide accurate data. Otherwise, the machine learning algorithm will not work. Apart from technical issues and data collection, educating the farmers is a recommendation for the future. If farmers understand the consequences of using incorrect fertilizer, then they will use technology to improve fertilizer use. Another recommendation for the future is a mobile app that shows NPK values of all areas. The mobile app should show the correct crop for the land as well. Furthermore, the app will show the inefficiency of incorrect crops. Such an app will encourage farmers to plant crops according to the NPK values of each land. However, researchers can only create such an app after extensive testing. Currently, no one has collected the NPK values in Pakistan. Lastly, other researchers should collaborate with the government to gather data on a massive scale. The data collection will improve the agriculture sector in Pakistan, and it will improve the economy of the country. So, ultimately it is the responsibility of the government to empower researchers in Pakistan. The government involvement is necessary for implementing recommendations in future.

8. Conclusion

We used machine learning to find the relationship between PK, and soil pH-CEC. The data set II was used for the nitrogen sample, and moisture- temperature was used to find the relationship as the data set II had no soil pH or CEC. We applied multiple machine learning algorithms for the samples. We found that multiple independent variables produce better output. The data set I had a small number of samples; therefore, the R square value was

low. Nonetheless, we identified a relationship. The Bayesian regression and ridge regression were the best methods in finding the relationship between pH-CEC, soil moisture-temperature and NPK.

Acknowledgments

I am sincerely thankful to NED University of Engineering & Technology for providing me with the opportunity to write a research paper as a result of Independent Study Project on the topic "Prediction of soil macro nutrients using machine learning algorithm". I can't express enough thanks to my supervisor Dr Muhammad Khurram for providing me with this opportunity & providing me continuous support & guidance. I offer my sincere appreciation to him for all that he has done for me. My completion of this project could not have been accomplished without my fellow members, Ahsan Ur Rehman & Dr Saad Qasim. To my family – thank you for allowing me time away from you to research & study. Finally, I am thankful to Mr. Mubashir Fazal for guiding me at the critical stage of this write up. Without his support it would have been very difficult for me to prepare the paper so meaningful & interesting. Through this study, I have learnt a lot about data science & its application in soil studies. I hope this research would help the other researchers to analyse the data in a very different way.

References

[1]. M. Sillanpää, Micronutrients and the Nutrient Status of Soils: A Global Study, Food & Agriculture Org, 1982.

[2]. M. Asaduzzaman and T. Asao, Improvement of Quality in Fruits and Vegetables Through Hydroponic Nutrient Management, BoD – Books on Demand, 2019.

[3]. H. Mirzakhaninafchi, I. M. Mishra and A. M. Nafchi, "Study on Soil Nitrogen and Electrical Conductivity Relationship for Site-," in 2017 ASABE Annual International Meeting, Washington, 2017.

[4]. L. G. f. C. S. T. a. P. Analysis, J. Benton Jones, Jr., CRC Press, 2001.

[5]. N. Tilley, "Fertilizer Numbers – What Is NPK," gardeningknowhow.com, 4 April 2018. [Online]. Available: https://www.gardeningknowhow.com/garden-how-to/soil-fertilizers/fertilizer-numbers-npk.htm. [Accessed 3 October 2019].

[6]. U. Hanif, "Pakistan's agriculture productivity among the lowest in the world," The Express Tribune, 24 January 2018. [Online]. Available: https://tribune.com.pk/story/1616347/2-pakistans-agriculture-productivity-among-lowest-world/. [Accessed 14 October 2019].

[7]. World Bank, "Collection of Development Indicators," World Bank, 2014.

[8]. Dawn, "Misuse of fertilisers reduces yield," Dawn, 4 February 2015. [Online]. Available: https://www.dawn.com/news/1161320. [Accessed 14 October 2019].

[9]. D. Li and Y. Chen, Computer and Computing Technologies in Agriculture: 5th IFIP TC 5, SIG 5.1 International Conference, CCTA 2011, Beijing, China, October 29-31, 2011, Proceedings, Part 1, Springer Science & Business Media, 2012.

[10]. K. Abhang, S. Chaughule, P. Chavan and S. Ganjave, "Soil Analysis and Crop Fertility Prediction," International Research Journal of Engineering and Technology, vol. V, no. 3, pp. 3106-3108, 2018.

[11]. D. Vadalia, M. Vaity, K. Tawate and D. Kapse, "Real Time soil fertility analyzer and crop prediction," International Research Journal of Engineering and Technology (IRJET), vol. IV, no. 3, pp. 1497-1499, 2017.

[12]. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

[13]. M.S.Suchithra and M. L.Pai, "Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters," 15 May 2019. [Online]. Available: https://doi.org/10.1016/j.inpa.2019.05.003. [Accessed 3 October 2019].

[14]. Enterprise Technology Review, "AI and Automation in Smart Farming," Enterprise Technology Review, 14 May 2019. [Online]. Available: https://www.enterprisetechnologyreview.com/news/ai-and-automation-in-smart-farming-nwid-196.html. [Accessed 13 October 2019].

[15]. Kellogg Biological Station, "Soil Test Lab Analyses - Agronomic," 7 February 2019. [Online]. Available: https://lter.kbs.msu.edu/datatables/52. [Accessed 6 October 2019].

[16]. Kellogg Biological Station, "Soil Inorganic Nitrogen, Moisture and Temperature," 4 December 2018. [Online]. Available: https://lter.kbs.msu.edu/datatables/148. [Accessed 6 October 2019].

[17]. D. A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, 2009.

[18]. A. K. M. E. Saleh, M. Arashi and B. M. G. Kibria, Theory of Ridge Regression Estimation with Applications, John Wiley & Sons, 2019.

[19]. scikit-learn, "Generalized Linear Models," scikit-learn, 4 October 2019. [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares. [Accessed 4 October 2019].

[20]. P. D. Hoff, A First Course in Bayesian Statistical Methods, Springer Science & Business Media, 2009.

[21]. S. L. Jackson, Statistics Plain and Simple, Cengage Learning, 2009.