Decision Tree Classification of Digital Soil, Weather, Crop Mapping and Yield Prediction using Linear Regression with region influences

Anna Rini, Hemalatha N and *Raji Sukumar
annarini008@gmail.com, rajivinod.a@gmail.com*, hemalatha@staloysius.ac.in
*Corresponding author

ABSTRACT

This project deals with the study of soil properties ,crop and the regional influences along with their dependencies which would be further used for a digital map. Both classification and regression algorithms were carried out and a decision tree as well as a decision regressor tree was plotted to finalise the results. Out of the 6 classification algorithms applied decision tree gave the highest accuracy of 95.24% and linear regression gave the best accurate results of 100% among the 3 regression algorithms.

Keywords: Machine learning, classification, accuracy, model creation, regression

1 Introduction

Economy is highly dependent on agriculture which is the basic building block of the economic system of a country.Along with providing of food and raw materials it is a source of various employment opportunities as well. Soil is an important feature when it comes to agriculture which provides sufficient nutrients for a successful cultivation.Likewise, weather plays an important role in agricultural production. Weather aberrations may cause physical damage to crops and erosion. The quality of crop produced during movement from field to storage and transport to plug depends on weather. Digital mapping is the process by which a set of knowledge is compiled and formatted into a reflection. The primary function of this technology is to supply maps that give accurate representations of a specific area, detailing major road arteries and other points of interest Digital mapping is now used world wide to map soil classes and its features. The introduction of machine learning (ML) algorithms in DSM is a game changing way for scientists in the production of maps. Machine learning is widely used to map soil properties or classes just like in other fields of study

This study is aimed at providing digital maps that would help us to get a clear idea about the soil, weather and crop of seven districts of Bangladesh along with yield prediction. Various Classification and regression algorithms was used for this analysis which are discussed in further sections.

2 Methodology

In this section we have discussed the dataset and machine learning techniques that were used for both classification and regression.

2.1 Dataset

The data set is collected from the Kaggle and is of 7 districts from Bangladesh. There were 150 data's in total. All the variables except two were in the categorical datatype while the rest were all numerical variables.

2.2 Features Used

Table 1 depicts the different features used and their data types.

Table 1: Features and their datatypes

Variable Name	Data Type
District	Categorical
Year	Numerical
Avg_rainfall	Numerical
Max_temperature	Numerical
min_temperature	Numerical
Aus	Numerical
Aman	Numerical
Boro	Numerical
Wheat	Numerical
Potato	Numerical
Jute	Numerical
Humidity	Numerical

Storm	Numerical
Urea	Numerical
Tsp	Numerical
Mp	Numerical
DAP	Numerical
inundationland_Highland	Numerical
inundationland_mediumhighland	Numerical
inundationland_lowland	Numerical
inundationland_mediumlowland	Numerical
inundationland_verylowland	Numerical
Miscellaneous Land	Numerical
Calcareous Alluvium	Numerical
Noncalcareous alluvium	Numerical
Acid basic clay	Numerical
Calcareous brown floodplain soil	Numerical
Calcareous grey floodplain soil	Numerical
Calcareous dark grey floodplain soil	Numerical
Noncalcareous frey floodplain soil	Numerical
Noncalcareous dark grey floodplain soil	Numerical
peat	Numerical
Made land	Numerical
Noncalcareous brown floodplain soil	Numerical
Shallow red brown terrace soil	Numerical
Deep red brown terrace soil	Numerical
Brown mottled terrace soil	Numerical
Shallow grey terrace soil	Numerical
Deep grey terrace soil	Numerical
Grey valley soil	Numerical
Brown hill soil	Numerical
Grey piedmont soil	Numerical

2.3 Work Flow

The characteristic of features used in this research work involved micro climatic conditions, soil properties and area of the field. Initial steps was data cleaning process which involved detection of missing values, skewness outliers and their treatments. Normalization techniques were applied for data transformations. By using Principal Component Analysis, data reduction was performed on the dataset. Further, data was converted into training and testing. After the completion of successive splitting of the data, different regression algorithms such as Linear Regression, Random forest Regressor and decision tree regression algorithms were applied in order to predict the potato yield and decision tree, Random Forest, SVM, Gaussian Naive Bayes, logistic regression algorithms for classification.

Based on the accuracies obtained from different models, the best algorithm was selected for the model deployment. The accuracies obtained from the different model was compared in order to find the best algorithm for model deployment. For entire research work python environment was used. Entire workflow is depicted in Figure 1.

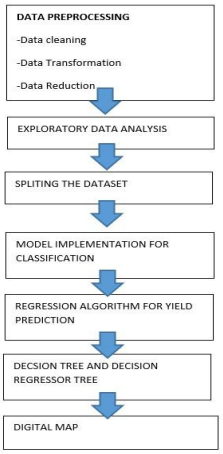

Fig 1: Workflow of the model

2.4 Pre-processing

2.4.1 Checking Missing Value

The missing values in the data affects the accuracy of the model. The problem with the missing data is, most of the algorithms do not accept the missing values. We cannot simply ignore the missing data from our dataset.

The first step of pre-processing is to check null values in the dataset. If there exists, missing data needs to be treated. There are different deletion and imputation techniques available in order to treat the missing values in the dataset. In this dataset we checked the missing value using isnull() command in python.

2.4.2 Checking Skewness

Skewness is the degree of asymmetry observed in a probability distribution. The skewness does not detect the outliers but it can give the direction to the outlier in the dataset. Using the skew() command in python we can easily detect the skewness existing in the data. The distplots of the different features in the dataset will give a visualization of

distributions in the dataset. Using different transformations we can treat the skewness in the data. Here we have used the box-cox transformation to treat the skewness.

2.4.3Outliers

Outliers are data points that are quite far away from the other data points. In other words, they are unusual values appearing in a dataset.. Outliers are an issue for many statistical analyses because they make tests to either miss significant findings or effect real results. Here it is found that there are outliers present in the yield. So here we plot the boxplots of all the variables in order to find the outliers and then used winsorizing method to treat them.

2.4.4 Data Transformation

Data transformation in machine learning models needs certain data processing operations.

1. Data cleansing – Removing superfluous and repeated data records from data will enhance the speed at which a business's machine learning model trains, also as improve their analysis.

2. Alter Data Types – Leveraging the right data types assists in saving memory usage and may be a requirement for predictions to be performed against it.

3. Convert Categorical Data to Numerical – There is a requirement for converting categorical variables to numerical variables in many of the ML analyses. This is yet another data transformation

2.5 Data Reduction

Since the dataset contains many features, we can perform dimensionality reduction techniques. For Dimensionality reduction we have used Principal Component Analysis(PCA) which converts the data from high dimension to lower dimension. This process will not lose much information as it holds most of the information in the data even after the dimensionality reduction using PCA.

2.6 Data Splitting

Here training and testing of the dataset is done The loaded data is divided into a training and a testing set with a division ratio of 80% or 20%, such as 0.8 or 0.2. During the learning phase, a classifier is used for formation of available input data. In this step, classifiers support data is and preconceptions to approximate and classify the function is created. Data is tested during the test phase The final data is formed during preprocessing and is processed by the machine learning module. So this process is certainly an essential step for any supervised machine learning or data science application.

2.7 Algorithms

2.7.1 Random forest

It is a supervised machine learning algorithm used for both classification and regression algorithm.eventhough it's mainly used for classification problems..Steps for this algorithm are:
Step 1: Selection of random samples
Step 2: Creation of decision tree and prediction of results form each tree
Step 3: Voting for each prediction result
Step 4: select the foremost voted prediction result

Random forest, consists of a number of decisions that works as an ensemble. Each decision tree gives out a prediction and the class with foremost votes becomes the prediction

2.7.2 KNN Classifier

The k-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm that can be used for both regression and classification. It is easy to implement and understand but is quite slow in working. It works by finding distance between the question and other data points in the dataset. Selecting the required number of closer points, and then voting for the most frequent or averages.

2.7.3Decision tree classifier

Decision is a supervised machine learning algorithm which is used for both classification and regression Here the internal nodes are representation of features and branches represent choice rules and every leaf node represents the results. Decision nodes will not make any decision and they have multiple branches while,leaf nodes are those with no long braches but are results of the decsions. Decsions are performed on the features of the dataset. It is a graphical representation of the possible outcomes of a problem within certain conditions

2.7.4Gaussian Naïve Bayes

Gaussian Naive Bayes is a variant of Naive Bayes following Gaussian distribution and supports continuous classification. Naive Bayes are a version of supervised machine learning classification algorithms using bayes theorem. It is a simple classification technique with high functionality

2.7.5Logistic Regression

Logistic regression is named after a function used within the method, the logistic function or sigmoid function $f(value)=1/(1+e^{-value})$.
Logistic regression uses this equation as the representation, Input values are combined linearly using coefficient values to predict an output value The output value being modelled is binary values (0 or 1) rather than a numeric value.

2.7.6 Support vector machines

(SVMs) are a group of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines are: •Effective in high dimensional spaces. • Effective in cases where number of dimensions is greater than the number of samples. • Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. • Versatile: Different Kernel functions can be specified for the decision function. Common kernels are provided, but it's also possible to specify custom kernels.

2.7.7Linear Regression

Linear Regression is a method of modelling a target value based on independent predictorsIt is an attractive model because the representation is so simple. Representation is a linear equation that combines a specific set of input values (x), the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output values are numeric.

3.Results and Discussions

In this section, we have described the results obtained for pre-processing as well as model accuracies for different algorithms for classification and regression.

Missing data detection and treatment

Here we had 32 missing value observations in the variable "area" (Figure 2). Since it was a numerical variable we could do mean imputation to tackle the issue.

Outlier detection

Figure 3 sows the outliers present in the data and Figure 4 depicts the treatment of the same using winsorizing method.

District	0
year	0
area	35
avg_rainfall	0
max_temperature	0
min_temperature	0
aus	0
aman	0
boro	0
wheat	0
potato	0
jute	0
humidity	0
storm	0
urea	0
tsp	0
mp	0
DAP	0
inundationland_Highland	0
inundationland_mediumhighland	0
inundationland_lowland	0
inundationland_mediumlowland	0
inundationland_verylowland	0
Miscellaneous Land	0
Calcareous Alluvium	0
Noncalcareous Alluvium	0
Acid Basin Clay	0
Calcareous Brown Floodplain Soil	0
Calcareous Grey Floodplain Soil	0
Calcareous Dark Grey Floodplain Soil	0
Noncalcareous Grey Floodplain Soil	0
Noncalcareous Dark Grey Floodplain Soil	0
Peat	0
Made-Land	0
Noncalcareous Brown Floodplain Soil	0
Shallow Red-Brown Terrace Soil	0
Deep Red-Brown Terrace Soil	0
Brown Mottled Terrace Soil	0
Shallow Grey Terrace Soil	0
Deep Grey Terrace Soil	0
Grey Valley Soil	0
Brown Hill Soil	0
Grey Piedmont Soil	0
soil moisture	0

District	0
year	0
area	0
avg_rainfall	0
max_temperature	0
min_temperature	0
aus	0
aman	0
boro	0
wheat	0
potato	0
jute	0
humidity	0
storm	0
urea	0
tsp	0
mp	0
DAP	0
inundationland_Highland	0
inundationland_mediumhighland	0
inundationland_lowland	0
inundationland_mediumlowland	0
inundationland_verylowland	0
Miscellaneous Land	0
Calcareous Alluvium	0
Noncalcareous Alluvium	0
Acid Basin Clay	0
Calcareous Brown Floodplain Soil	0
Calcareous Grey Floodplain Soil	0
Calcareous Dark Grey Floodplain Soil	0
Noncalcareous Grey Floodplain Soil	0
Noncalcareous Dark Grey Floodplain Soil	0
Peat	0
Made-Land	0
Noncalcareous Brown Floodplain Soil	0
Shallow Red-Brown Terrace Soil	0
Deep Red-Brown Terrace Soil	0
Brown Mottled Terrace Soil	0
Shallow Grey Terrace Soil	0
Deep Grey Terrace Soil	0
Grey Valley Soil	0
Brown Hill Soil	0
Grey Piedmont Soil	0
soil moisture	0

dtype: int64

Fig 2: Checking Missing values

Fig 3: Showing outliers

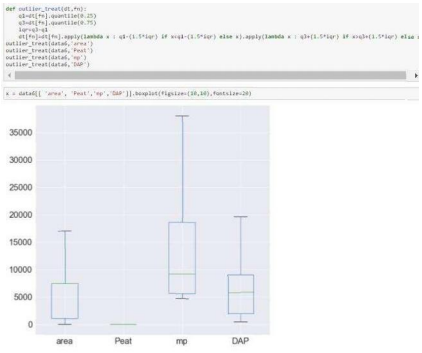

Fig 4: Outlier treatment using winsorizing method

Skewness in the data

Figure 5 shows that there exists skewnesss in the data. So we used box-cox transformation to treat this skewness. The resultant skewness after the transformation is given in Figure 6.

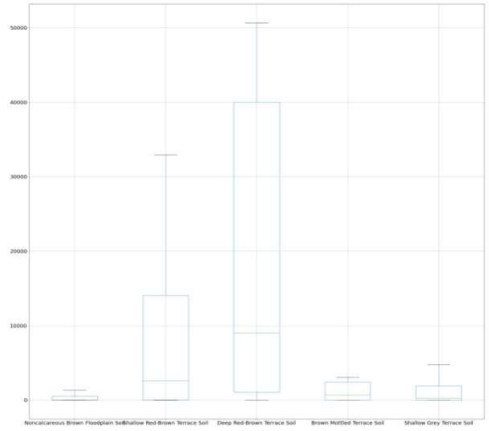

Fig 5: Checking the skewness in data

```
skw =data6.skew(axis=0,skipna=True)
print(skw)
```

```
year                                    0.000000
area                                    0.692757
avg_rainfall                           -0.336988
max_temperature                        -0.645450
min_temperature                         1.402460
aus                                     1.661500
aman                                    1.186077
boro                                    0.926972
potato                                  0.323992
humidity                               -0.172384
urea                                    0.777684
tsp                                     1.690496
mp                                      1.327272
DAP                                     1.220174
inundationland_Highland                 0.868909
inundationland_mediumhighland           0.827482
inundationland_lowland                  0.508833
inundationland_mediumlowland            1.445530
inundationland_verylowland              2.086215
Miscellaneous Land                      0.483701
Calcareous Alluvium                     2.086342
Noncalcareous Alluvium                  1.370435
Acid Basin Clay                         0.764330
Calcareous Brown Floodplain Soil        2.086222
Calcareous Grey Floodplain Soil         2.086220
Calcareous Dark Grey Floodplain Soil    2.086715
Noncalcareous Grey Floodplain Soil      0.063875
Noncalcareous Dark Grey Floodplain Soil 1.350130
Peat                                    0.000000
Made-Land                               2.065453
Noncalcareous Brown Floodplain Soil     2.194603
Shallow Red-Brown Terrace Soil          1.063136
Deep Red-Brown Terrace Soil             0.508681
Brown Mottled Terrace Soil              0.290027
Shallow Grey Terrace Soil               1.862481
Deep Grey Terrace Soil                  1.235975
Grey Valley Soil                        0.913023
Brown Hill Soil                         2.158593
Grey Piedmont Soil                      2.233135
soil moisture                           0.566333
District                                0.000000
storm                                  -0.609422
dtype: float64
```

Fig 6: After box-cox transformation

Visualization results

Heatmap in Figure 7 indicates the correlation between the variables. Some of the variables are positively correlated and some them are negatively correlated and variable aus has a high correlation with the area. Barchart is plotted in Figure 8 and 9 for all variables. Figure 10 depicts the pie chart.

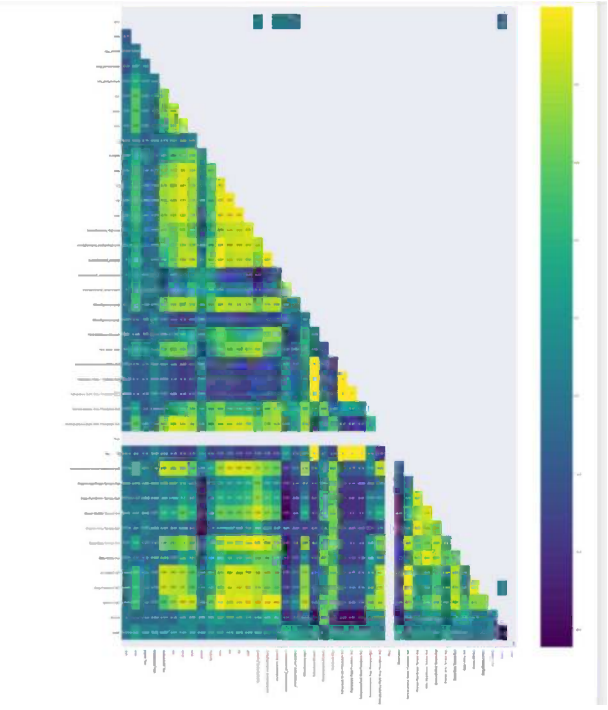

Fig 7: Heatmap

Bar Charts

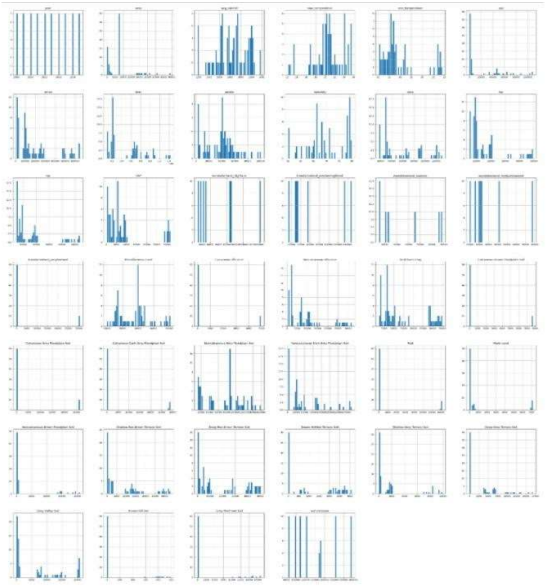

Fig 8: Barcharts of all the variablesles

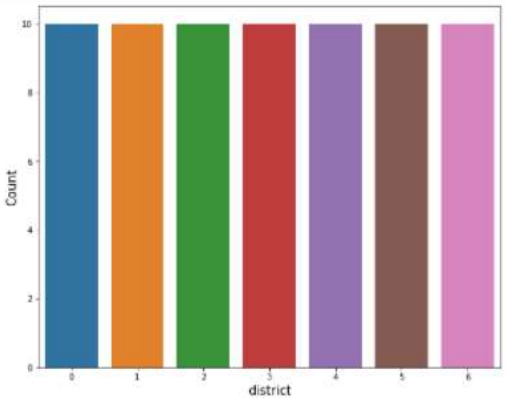

Figure 9Barcharts of district variable

Algorithm	accuracy
Decision tree	95.24
Random forest	92.857
Knn	92.857
Gaussian nb	92.857
Svm	85.714
Logistic regression	92.857

```
District_count=[10,10,10,10,10,10,10]
District_labels=['1','2','3','4','5','6','0']
plt.pie(District_count,labels=District_labels,radius=2,autopct='%0.1f%%',shadow=True)
```

```
([<matplotlib.patches.Wedge at 0x5d21868>,
  <matplotlib.patches.Wedge at 0x4c9bfd0>,
  <matplotlib.patches.Wedge at 0x4c9b658>,
  <matplotlib.patches.Wedge at 0x4c9bee0>,
  <matplotlib.patches.Wedge at 0x5b7ae50>,
  <matplotlib.patches.Wedge at 0x5b7ab80>,
  <matplotlib.patches.Wedge at 0x5b7a310>],
 [Text(1.982131490234419, 0.9545442658258875, '1'),
  Text(0.4895459256088259, 2.1448414362651143, '2'),
  Text(-1.371677736632888, 1.7200291238306693, '3'),
  Text(-2.1999999999999784, -3.8896829786607e-07, '4'),
  Text(-1.3716772535105801, -1.7200295091078155, '5'),
  Text(0.4895465280524509?, -2.144841298761237, '6'),
  Text(1.982131758346992 4, -0.9545437090842205, '0')],
 [Text(1.0811626310369555, 0.5206605086123022, '14.3%'),
  Text(0.2670250503320868 4, 1.1699135106900622, '14.3%'),
  Text(-0.7481878563452072, 0.9381977039076377, '14.3%'),
  Text(-1.1999999999999882, -1.6852816247817854e-07, '14.3%'),
  Text(-0.7481875928239526, -0.9381979140588884, '14.3%'),
  Text(0.2670253789377005, -1.1699134356879475, '14.3%'),
  Text(1.0811627772801775, -0.5206602049550293, '14.3%')])
```

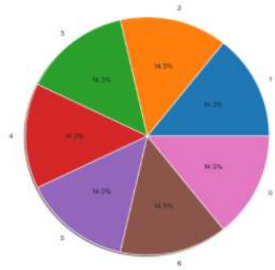

Figure 10: Pie chart

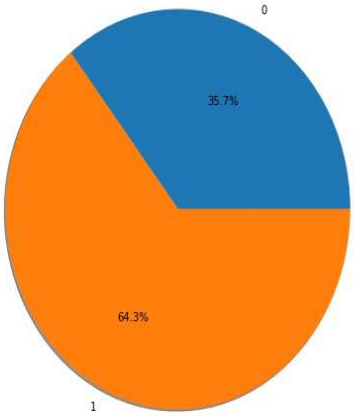

Fig 11: Output class conversion and checking imbalance in the data

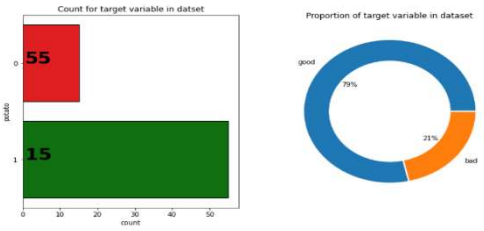

Fig 12: Treatment using smote

Figure 11 show the class imbalance present in the dataset which are treated by SMOTE and result is showed in Figure 12.

Algorithms:

Result obtained for different classification algorithms are as follows: For decision tree classifier had an accuracy of 95.24% was obtained, Random forest classifier, Gaussian Nb, K nearest neighbors classifier and Logistic regression obtained 92.8 %. For Support vector machine linear kernel obtained 92.8 and RBF kernel and poly kernel obtained 85.7%. All these results are depicted in Table 2.

Table 2:Accuracy results of classification

Results obtained for regressors are as follows: Linear regression obtained an accuracy of 100%, Random forest 94.1% and Decision tree 94.2%. The structure chart of decision tree regressor is depicted in Figure 13.

Table 3:accuracy results of regression

alogorithm	accuracy
Linear regression	100
Random forest regressor	94.19
Decision tree regressor	94.926

From the classification and regression results we can conclude that among classifiers Decision tree and KNN are the best with 100% and in regressors Random forest is the best with 100%.

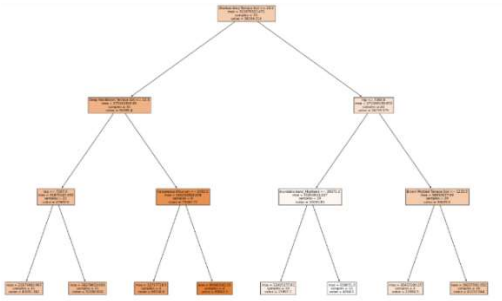

Fig 13 Final chart of decision tree regressor

4.CONCLUSIONS

From the analysis it is clear that decision tree is the best classifier for classification of the yield and linear regression is the accurate regressor for predicting the yield. From the decision tree and decision regressor tree it is visible that the variables,storm,and shallow deep terrace soil are the most contributing features to the yield class.

5. ACKNOWLEDGEMENT

The authors extend their appreciation to the Deputyship of RESEARCH AND INNOVATION wing of TECHTERN Pvt. Ltd. through SMART-AGRO research and for providing all support for this research work with the project number TTRD-DS-03-2021. This is also an extension of a post-doctoral research program under Kannur University.

REFRENCES

[1] Taghizadeh-Mehrjardi, R., *et al*., "Digital mapping of soil classes using decision tree and auxiliary data in the Ardakan region, Iran." *Arid Land Research and Management* 28.2 (2014): 147-168.

[2] Wadoux, Alexandre MJ-C., Budiman Minasny, and Alex B. McBratney. "Machine learning for digital soil mapping: applications, challenges and suggested solutions." *Earth-Science Reviews* (2020): 103359.

[3] Giasson, Elvio, *et al*., "Decision trees for digital soil mapping on subtropical basaltic steeplands." *Scientia Agricola* 68 (2011): 167-174.

[4] Taghizadeh-Mehrjardi, Ruhollah, et al. "Digital mapping of soil classes using ensemble of models in Isfahan region, Iran." *Soil Systems* 3.2 (2019): 37.