International Journal of
Science and Engineering Investigations
vol. 8, issue 92, September 2019
Received on September 24, 2019
ISSN: 2251-8843

A Study on Data Characteristics of Some Wheat Varieties for Machine Learning

Eray Onler[1], İlker H. Celen[2]
[1,2]University of Namik Kemal, Faculty of Agriculture, Dept. of Biosystem Engineering, 59030-Suleymanpasa-Tekirdag, Turkey
([1]erayonler@nku.edu.tr, [2]icelen@nku.edu.tr)

Abstract- Wheat (Triticum spp.) is a grass that is cultivated worldwide. It is the most important human food grain and ranks second in total production as a cereal crop behind maize. Correct classification of wheat varieties is giving chance to have consistent processing and end product performance. Today most of the classification efforts are doing manually. Machine learning gives chance to us for automating this process. In machine learning approach instead of implicitly programming all the steps, we collect the data and computer will find structure in the data to classify wheat by using statistical learning methods.

Machine learning is about learning and extracting some properties of a data (it is wheat seed measurements in this case) and applying them to new data (unseen data). This is why a common practice to split the data into two sets in machine learning model evaluation. We call the training set on which we learn data properties and we call the testing set on which we test these properties. However in any case machine learning model's prediction accuracy depends on the quality of data that it fed.

In general, a learning problem considers a set of n samples of data and m features about these samples and then tries to predict properties of unknown data.

In this study we used dataset which has collected in storage of seeds obtained fields on Namik Kemal University. This dataset consist seven different physical measurements (area A, perimeter P, compactness C = 4 x pi x A/P^2, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove.) of three different varieties of wheat (Kama, Rosa, Canadian).

Python 2.7.11, sklearn 0.18, pandas 0.19.0 and seaborn 0.7.1 version were used in the study. Dataset was randomly splitted to train and test splits. Train data consists 0.67 and test data consists 0.33 of total data. Missing datas are inputted with the column means. Decision Tree Classifier were used with following settings; criterion: gini, max_depth: 8 and max_fetures: None. Most important feature is groove with 56% related to model. Model prediction accuracy is %94 on test dataset.

Keywords*-Machine Learninig, Wheat, Decision Tree, Classification, Kama, Rosa, Canadian*

I. Introduction

Wheat (Triticum spp.) is a grass that is cultivated worldwide. It is the most important human food grain and ranks second in total production as a cereal crop behind maize. Correct classification of wheat varieties is giving chance to have consistent processing and end product performance. Today most of the classification efforts are doing manually. Machine learning gives chance to us for automating this process. In machine learning approach instead of implicitly programming all the steps, we collect the data and computer will find structure in the data to classify wheat by using statistical learning methods.

Decision trees are a supervised, probabilistic, machine learning classifier that are often used as decision support tools. Like any other classifier, they are capable of predicting the label of a sample, and the way they do this is by examining the probabilistic outcomes of your samples' features. Decision trees are named after their tree graph structure.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression) (Friedl and Brodley, 1997)

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and outpt variables. A decision tree algorithm is a process that split data set into subsets recursively. First it is dividing entire population or sample into two or more homogeneous sets and this is called root node. Every splitting nodes are called as decision node and the final nodes are called as terminal node or leaf. Decision tree output is very easy to understand. Also it is good way to see importance of different variables.

Wheat (Triticum spp.) is a major food source in the world and is cultivated worldwide. It is the most important human food grain and ranks second in total production as a cereal crop behind maize. Wheat grain is the main ingredients of breads, cookies, cakes, pasta, noodles etc. and also it is being used for fermentation of beer, alcohol, vodka or biofuel. Even the side product like bran have economic value.

Identification of varieties mostly based on labor force currently. However this is not the most efficient way because expert judgment can be problematic due to experience and work load. Because of wrong classification economic losses can be occurred. Autonomous classification systems can increase the efficiency (Olgun et al., 2016).

In this study, using dataset which has collected in storage of seeds obtained fields on Namik Kemal University, classification of some wheat seeds classified by Decision Tree Classifier. This dataset consist seven different physical measurements of three different varieties of wheat (Kama, Rosa, Canadian).

II. MATERIAL AND METHOD

In this study we used dataset which has collected in storage of seeds obtained fields on Namik Kemal University. This dataset consist seven different physical measurements (area A, perimeter P, compactness C = 4*pi*A/P^2, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove.) of three different varieties of wheat (Kama, Rosa, Canadian).

Python 2.7.11, sklearn 0.18, pandas 0.19.0 and seaborn 0.7.1 version were used in the study. Dataset was randomly splitted to train and test splits. Train data consists 0.67 and test data consists 0.33 of total data. Missing datas are imputed with the column means. Decision Tree Classifier were used with following settings; *criterion: gini, max_depth: 8 and max_fetures: None*.

The first step was reading our data as pandas dataframe structure. Dataframe is special Pandas data type for storing information as tabular data in row and columns. All rows are indicating samples and columns indicates features of this sample. We used *read_csv()* function of pandas for this purpose.

data = pd.read_csv("wheat.data.txt")

The next step is exploratory data analysis (EDA). EDA makes possible to understand our data structure and what features it have, that we can use to predict wheat type. We can divide exploratory data analysis into two parts, numerical and visual.

It is always good to check general structure of data with pandas *head()* function (Table 1). This gives us a broader view of our data. As we noticed we have 8 features (id, area, perimeter, compactness, length, width, asymmetry, groove) and 1 outcome (wheat type). ID feature can't be use for prediction of wheat type. Because it is just showing and order but does not contain any information about wheat type. So we dropped id column by using *drop()* function.

data.drop("id", axis=1, inplace=True)
data.head()

After that we separate our independent variables (predictive features) and dependent variable (outcome feature).

X = data[["area", "perimeter", "compactness", "length", "width", "asymmetry", "groove"]]
y = data[["wheat_type"]]

TABLE I. GENERAL STRUCTURE DATA

	area	perimeter	compactness	length	width	asymmetry	groove	wheat_ type
0	15.26	14.84	0.8710	5.763	3.312	2.221	5.220	kama
1	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	kama
2	14.29	14.09	0.9050	5.291	3.337	2.699	4.825	kama
3	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	kama
4	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	kama

It is always good practice to check the shape and size of our independent and dependent variables. We did this by following python code.

print("X shape is : {}".format(X.shape))
print("y shape is : {}".format(y.shape))

Our independent variables matrix have 210 rows and 7 columns and dependent variable vector have 210 rows and 1 column. When we checked how many different wheat type classes we have, we see that three; kama, canadian and rosa respectively.

As we see, wheat types are in string format. But machine learning models need numerical values. Because of this we needed to translate these string values to numerical category values. After numerical encoding 0, 1 and 2 assigned to wheat types canadian, kama and rosa respectively.

y["wheat_type"]=
y["wheat_type"].astype("category").cat.codes

After we encode wheat types it is good to check distribution of classes with *value_counts()* function. Our classes have nearly same distribution which is good. If our classes had uneven distribution we must apply stratified methods for dividing into train and test sets.

Canadian	0.361905
Rosa	0.323810
Kama	0.314286

Machine learning is about learning and extracting some properties of a data (it is wheat kernels measurements in this case) and applying them to new data (unseen data) (Pedregosa et al., 2011). Because of that reason we usually divide our dataset into two parts, training and test. We are creating our model on training set and then evaluating our model performance on test set. Our full dataset divided in random way to train (67%) and test (33%) set by using *train_test_split()* function.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=7)
```

After dividing our dataset we checked descriptive statistics of our training and test datasets. We noticed that compactness, width and groove in train dataset and also compactness and groove in test dataset have missing values. Most of the machine learning models cannot manage missing values so we need to treat them. There are many different approaches for treating missing values, like dropping them or filling with mean or median value of the related column. Dropping sample is not best choice because it also cause information lose. So we used filling by mean, because spread of compactness, width and groove features is not high.

```
X_train["compactness"].fillna(X_train["compactness"].mean()
, inplace=True)
```

```
X_train["width"].fillna(X_train["width"].mean(),
inplace=True)
```

```
X_train["groove"].fillna(X_train["groove"].mean(),
inplace=True)
```

After this point we examined our training dataset by visualizing. Test dataset is theoretically our unseen data and for preventing data leakage from test dataset to model we didn't take it into account for further analysis. We used histogram, boxplot and swarmplot to see the distribution, spread of our data and structure based on wheat classes. Also we checked relationships between variables by pairplot.

We used decision tree classifier for classifying purposes. Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Decision tree models are simple to understand and to interpret and result can be visualized. Different than other techniques it needs little data preparation. Most of the times only missing value treatment is enough. Decision tree models able to handle both numerical categorical data and also multi-output problems.

Before using decision tree classifier by scikit-learn we have setted parameters below:

criterion By default, Scikit learn uses Gini, which is an impurity rating. Alternatively, you could also make use of information gain, or entropy instead. We used Gini for this study.

splitter Lets you control of the algorithm chooses the best split or not. We'll discuss why that's importance once you move to random forest classifier. We used best option for this paramater.

max_features One of the possible splitter options for splitter above is called 'best'. scikit-learn runs a bunch of tests on your features to figure out which mechanism should be used when searching for the best split. This parameter limits the number of features to consider while doing this. We had just seven independent variables so we didn't restrict our max features and we took all them into our model.

III. RESULTS AND DISCUSSION

After dividing our complete data into train and test dataset we checked descriptive statistics of our seven independent variables (Table 2). There were missing values at compactness and width variables. Scales of variables are different but because we are using decision tree classifier this is not important at all. Area has the highest standard deviation. However we made visualization to understand better the distribution of our variables.

TABLE II. DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES

	area	perimeter	compactness	length	width	asymmetry
count	140.00	140.00	139.00	140.00	139.00	140.00
mean	14.63	14.46	0.87	5.56	3.23	3.74
std	2.81	1.26	0.024	0.58	0.37	1.47
min	10.59	12.41	0.81	0.90	2.64	0.85
25%	12.19	13.39	0.86	5.23	2.93	2.67
50%	14.22	14.27	0.87	5.49	3.19	3.65
75%	16.78	15.48	0.89	5.92	3.54	4.79
max	20.88	17.23	0.92	6.58	4.03	8.46

For first step of visual inspection we started with histogram and density plot to see spread of independent variables (Figure 1). Histogram is good to explain this but it depends on chosen bin size very much and we also add density plot to our graph to see this difference. As we can see from graphics none of the independent variables have normal distribution.

We noticed from histogram that all of our variables far from normal distribution.

For the second step we inspected relationship between independent variable by using pandas *corr()* function (Table 3). This table make it easier to see that some independent variables have strong linear relationships with each other. Ex. Area-perimeter, length-width, area-groove etc.

For the third step we inspected independent variables based on wheat classes by using seaborn *boxplot()* function (Figure 2.). Boxplot is good choice to check spread of data points, detection of outlier and difference between classes.

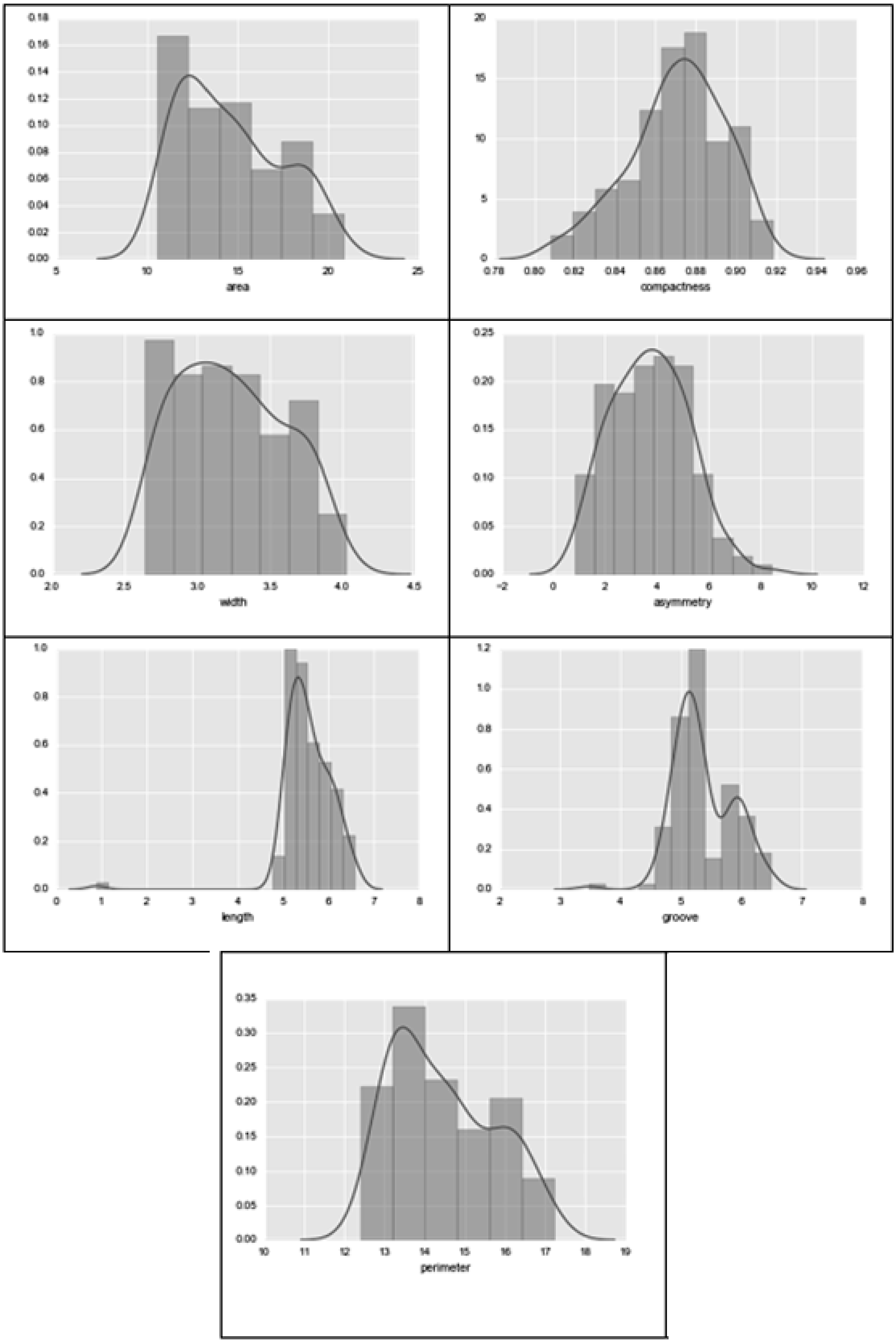

Figure 1. Histogram and density plots of independent variables.

TABLE III. CORRELATION TABLE OF INDEPENDENT VARIABLES

	area	perimeter	compactness	length	widht	asymmetry	groove
area	1.00	0.99	0.61	0.66	0.96	-0.26	0.80
perimeter	0.99	1.00	0.53	0.68	0.94	-0.25	0.83
compactness	0.61	0.53	1.00	0.27	0.76	-0.31	0.25
length	0.66	0.68	0.27	1.00	0.63	-0.24	0.86
width	0.96	0.94	0.76	0.63	1.00	-0.29	0.72
asymmetry	-0.26	-0.25	-0.31	-0.24	-0.29	1.00	-0.11
groove	0.80	0.83	0.24	0.86	0.71	-0.11	1.00

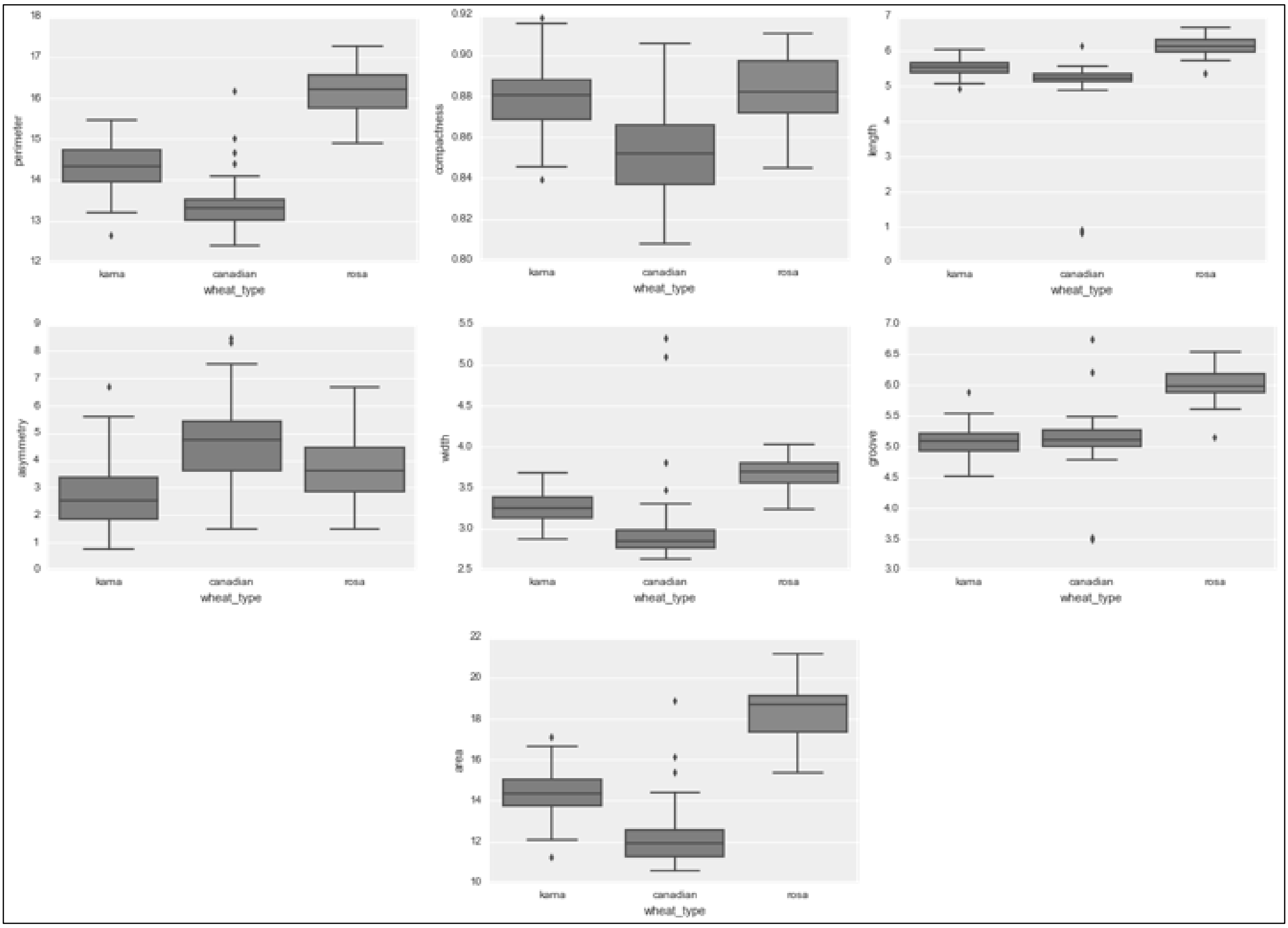

Figure 2. Boxplots of independent variables

After creating our model we applied this model onto our test dataset. Result classification accuracy was 94%. When we checked feature importance at classification process on test dataset we obtained following table (Table 4).

Decision tree model can be seen below. Groove is at the root because it is most important variable in the model (Figure 3).

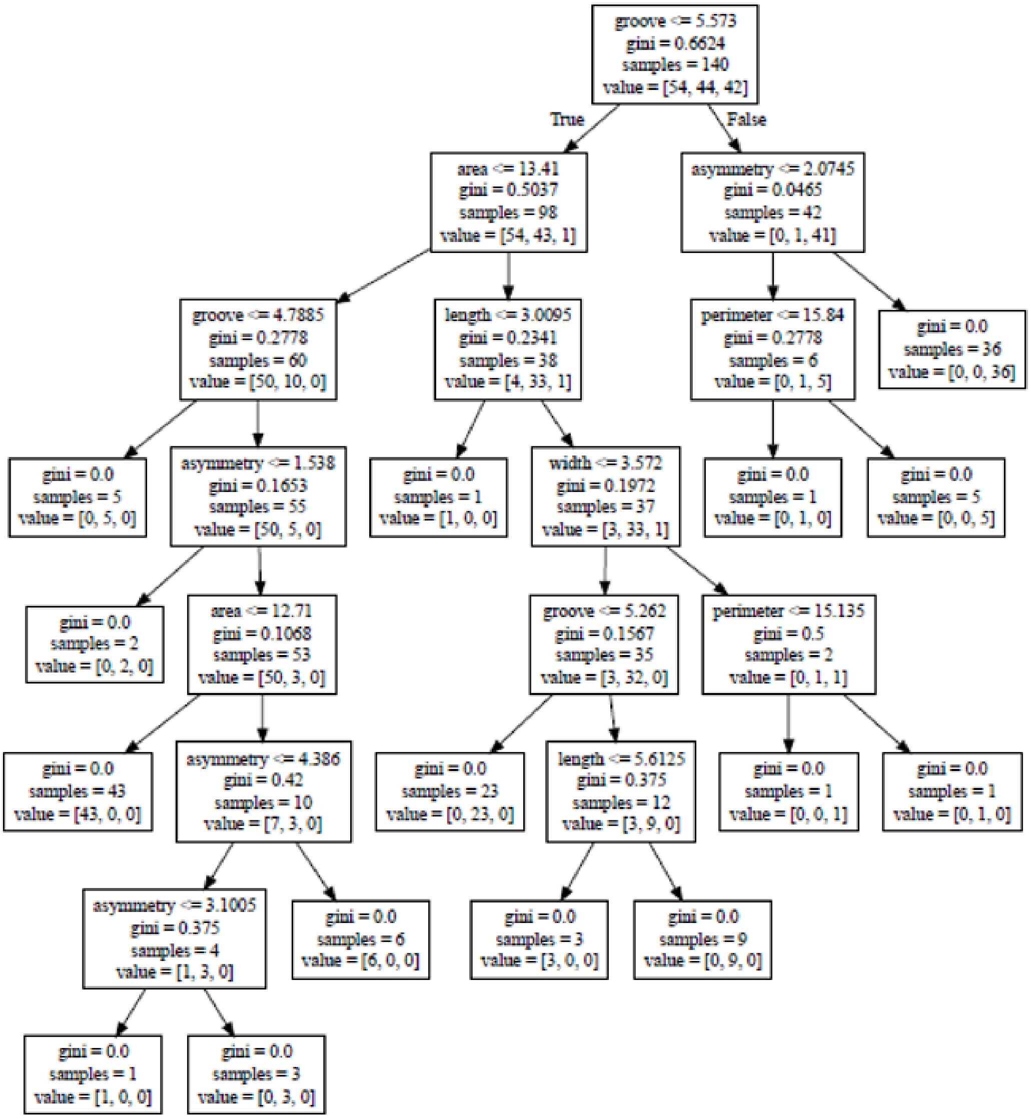

Figure 3. Decision tree model for classification.

TABLE IV. FEATURES IMPORTANCE OF THE MODEL

area	perimeter	compactness	length	width	asymetry	groove
0.27243414	0.02875334	0.00	0.06574565	0.0087509	0.05535693	0.56895904

IV. CONCLUSION

We used measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each randomly selected for creating a classification model. We applied machine learning process from reading and preparing the data to evaluate the model test dataset. This dataset is examined with exploratory data analysis first and then dataset divided into test and train datasets. Classification model is trained on train dataset and obtained model, applied on test dataset to see the accuracy of model on unseen data.

For increasing the generalizability of the model amount of the instances at dataset can be increased. Also for reducing the model variance and increasing generalizability k-fold validation can be applied on training data.

As a result, wheat classification was done using machine learning system and wheat characteristics. Results was promisive to change human intervention by using supervised machine learning.

REFERENCES

[1] M. Charytanowicz, Niewczas J., Kulczycki P., Kowalski and P.A., Lukasik, S., & Zak, S., “A complete gradient clustering algorithm for features analysis of X-ray images. Information Technologies in Biomedicine”, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2011, pp. 15-24.

[2] M.A. Friedl, and Brodley, E., “Decision tree classification of land cover from remotely sensed data”, Remote Sensing of Environment, 1997, Volume 61: 3.

[3] M. Olgun, Onarcan A.O., Ozkan K., Isik S., Sezer O., OzgiSi K., Ayter N.G., Basciftci Z.B., Ardic M. and Koyuncu, O., “Wheat grain classification by using dense SIFT features with SVM classifier”, Computers and Electronics in Agriculture, Volume 122, March, Pages 185–190, 2016.

[4] F. Pedregosa, Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer, P., Weiss, R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay, E., “Scikit-learn: machine learning in Python”, Journal of Machine Learning Research, 12,2825-2830, 2011.

Eray Onler was born in 1984 in Istanbul, Turkey and graduated from the Department of Electrical & Electronics Engineering, Faculty of Engineering, Sakarya University in 2006. I completed my master's degree in 2013 and my PhD in 2018. In 2010, I was appointed to Tekirdag Namik Kemal University as a research assistant. My research areas are agricultural machines, plant protection machines and precision agriculture. I have carried out studies, patents on those fields. I was researcher at 8 completed projects. I have 3 book chapters and 17 published articles.

How to Cite this Article:

Onler, E. & Celen, I. H. (2019) A Study on Data Characteristics of Some Wheat Varieties for Machine Learning. International Journal of Science and Engineering Investigations (IJSEI), 8(92), 142-148. http://www.ijsei.com/papers/ijsei-89219-19.pdf