Problem Statement : Classify 15 types of Soybeans using Support Vector Machine Classifier and present it's accuracy.

Abstract :

From the last decade, machine learning models are being deployed in various industries in the market. Out of those industries, the agricultural industry is one of them. Soybeans industry is one of the industries in the agro-based market. Soybeans come in various shapes and sizes, some of which are rather difficult to distinguish from one another. To make this segregation process easier, we suggest a Support Vector Machine classifier to process soybean data which consists of attributes of 15 soybean classes and present it's accuracy on another test data which consist of another set of the same 15 classes.

H/W and S/W requirements :

H/W requirements -

Intel Core i7 5600u vPro
8 GB RAM
256 GB Hard Drive

S/W requirements -

OS - 64 bit, Windows 10 Pro
Anaconda Python 5.2.0
Libraries -

1. Scikit-learn
2. Seaborn
3. matplotlib

Introduction :

Classification :

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

Handling Missing values:

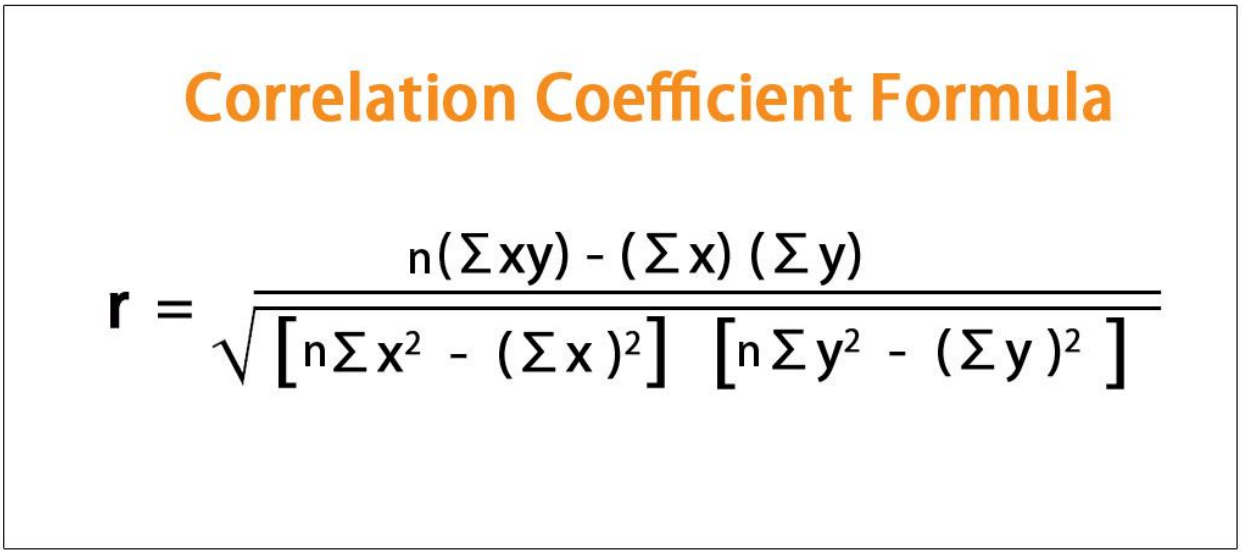
1. Drop the Columns with missing values
2. Imputation :

    Imputation fills in the missing value with some number. The imputed value won't be exactly right in most cases, but it usually gives more accurate models than dropping the column entirely.

    The default behavior fills in the mean value for imputation. Statisticians have researched more complex strategies, but those complex strategies typically give no benefit once you plug the results into sophisticated machine learning models.

Handling highly correlated data:

Correlation is the metric of dependency on one data column to other. Mathematically, correlation is defined as follows



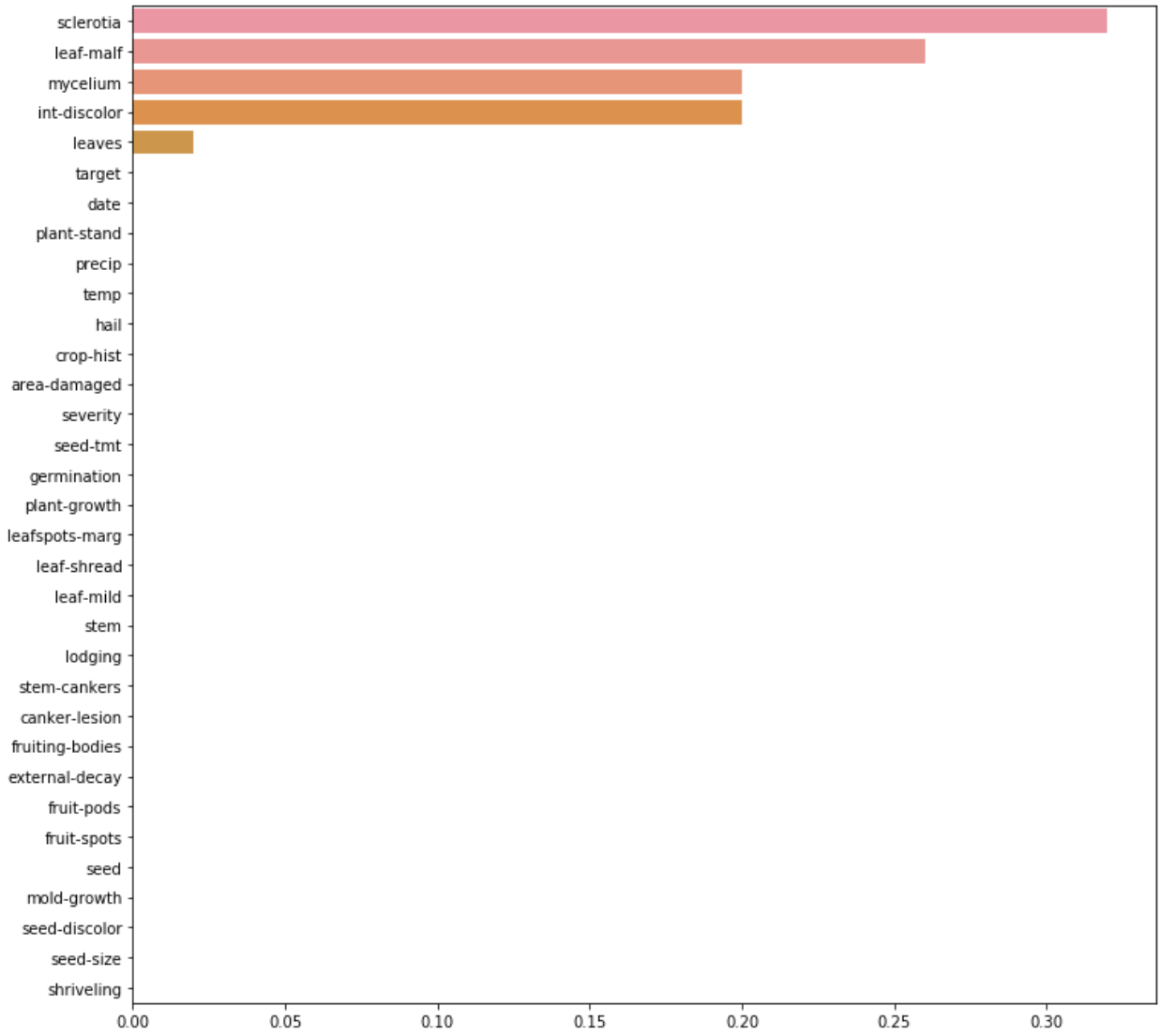
where x and y are attributes of columns X, Y

If some of the columns are highly correlated, they give the same information to the prediction of the target variable. Thus, if we remove one of the column in a highly correlated pair, the machine learning model will have a decreased amount of load while training as well as testing. In our data, we are removing one column from each pair of columns having correlation greater than 0.75.

Feature Selection using Feature Importances:

Ensemble techniques used in sklearn libraries provide a special output which provides importances of features in predicting target variables. These importances can be thresholded to

output the most relevant columns required for predicting target variables. The output of importance of features is given as the following graph.



Support Vector Machines:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

By guess work, we can achieve 1/15 ~ 6.6 accuracy on given target classes. We introduce a SVM classifier model which will achieve accuracy far greater than guess work, in predicting the target class variable.

Objective:

- Handle Missing values.
- Perform correlation analysis.
- Use ensemble techniques for feature selection.
- Build a classification model for prediction of disease in the Soybean Plants.
- Analyze the models using various classification metrics.

Scope:

- A classification system which is trained on 307 rows of data and tested on 387 rows of data
- System which can classify, given the required attributes, the soybean into one of 15 classes.

Test Cases:

1. Case 1: Without feature selection technique

	precision	recall	f1-score	support
0	0.00	0.00	0.00	51
1	0.43	0.62	0.51	24
2	0.00	0.00	0.00	10
3	0.00	0.00	0.00	10
4	0.00	0.00	0.00	52
5	1.00	0.71	0.83	24
6	1.00	1.00	1.00	10
7	0.00	0.00	0.00	10
8	0.00	0.00	0.00	10
9	0.23	1.00	0.37	51
10	0.50	0.60	0.55	10
11	0.00	0.00	0.00	4
12	0.00	0.00	0.00	10
13	0.00	0.00	0.00	10
14	0.00	0.00	0.00	10
accuracy			0.33	296
macro avg	0.21	0.26	0.22	296
weighted avg	0.21	0.33	0.23	296

2. Case 2: With feature selection technique.

	precision	recall	f1-score	support
0	0.93	0.80	0.86	51
1	1.00	0.83	0.91	24
2	0.00	0.00	0.00	10
3	0.00	0.00	0.00	10
4	0.39	0.98	0.56	52
5	0.94	0.71	0.81	24
6	1.00	1.00	1.00	10
7	1.00	1.00	1.00	10
8	1.00	0.40	0.57	10
9	0.95	0.73	0.82	51
10	0.00	0.00	0.00	10
11	1.00	1.00	1.00	4
12	0.00	0.00	0.00	10
13	1.00	0.70	0.82	10
14	1.00	1.00	1.00	10
accuracy			0.71	296
macro avg	0.68	0.61	0.62	296
weighted avg	0.73	0.71	0.69	296

Results:

- Accuracy without feature selection using feature importances of ensemble techniques by using SVC : 71 %
- Accuracy with feature selection using feature importances of ensemble techniques using SVC : 33 %
- Final output : SVC classifier without feature selection technique.

Conclusion: We have successfully interpreted the given data and designed and built a SVM classifier around it.

References:

1. https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
2. https://scikit-learn.org/
3. https://matplotlib.org/