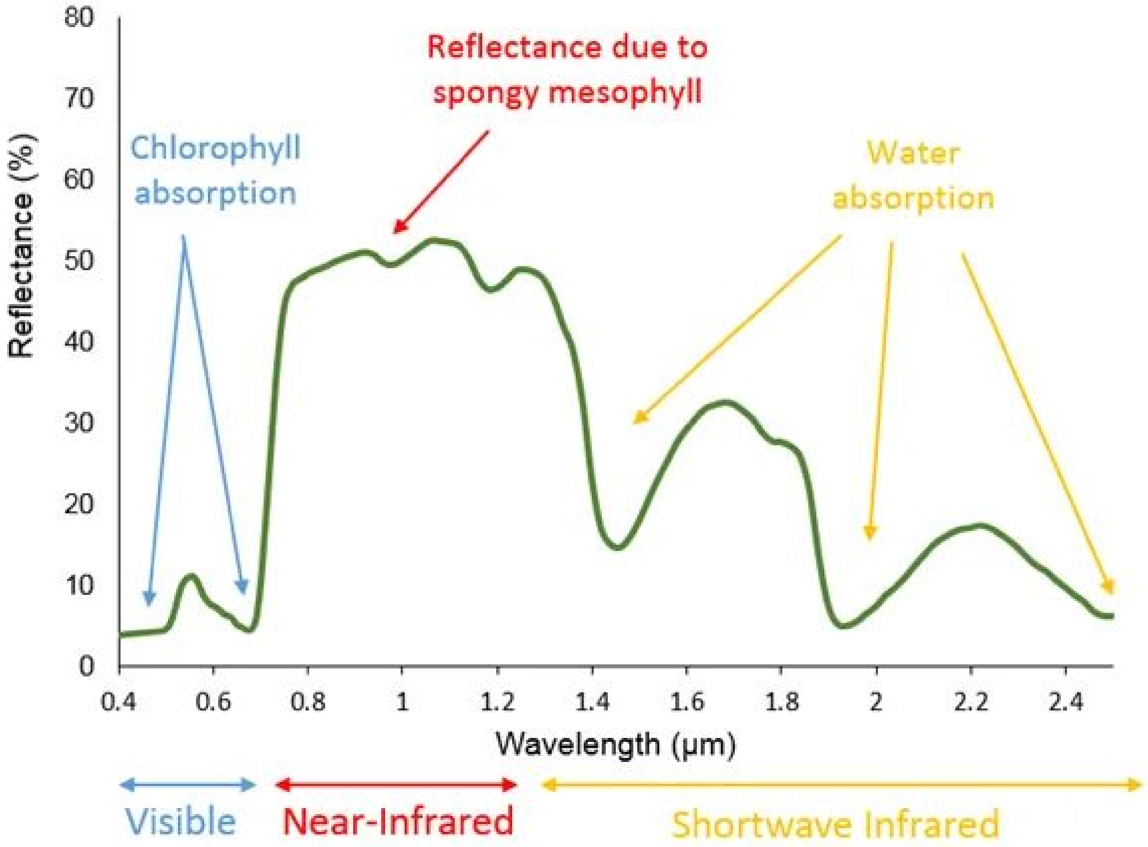
# REPORT

***Instructor*** *: Dr. KV Kale Sir (CS and IT Department)*

*Dr. Babasaheb Ambedkar Marathwada University*



**Onkar Sabnis**

(Department of Chemical Engineering)

IIT Kharagpur

## INTRODUCTION

The following report contains a brief overview of the project and research work carried out during winter research internship under the guidance of **Dr. KV Kale,Professor(Department of CS and IT)**, Dr. Babasaheb Ambedkar Marathwada University.

## HYPERSPECTRAL DATA ANALYSIS

**Hyperspectral non-imaging data** provides the spectral range from **400–2500nm** which has the ability to identify each and every unique materials on the surface. The plant species identification is critical task manually and computationally. So, **non-imaging hyperspectral data** is used over here for analysis.

## VEGETATION INDICES AND REMOTE SENSING(RS)

**Vegetation Indices (VIs)** are combinations of surface reflectance at two or more wavelengths designed to highlight a particular property of vegetation. They are derived using the reflectance properties of vegetation.

**Vegetation Indices (VIs)** obtained from remote sensing based canopies are quite simple and effective algorithms for **quantitative and qualitative evaluations** of vegetation cover, vigor, and growth dynamics, among other applications. These indices have been widely implemented within RS applications using different airborne and satellite platforms with recent advances using Unmanned Aerial Vehicles (UAV)

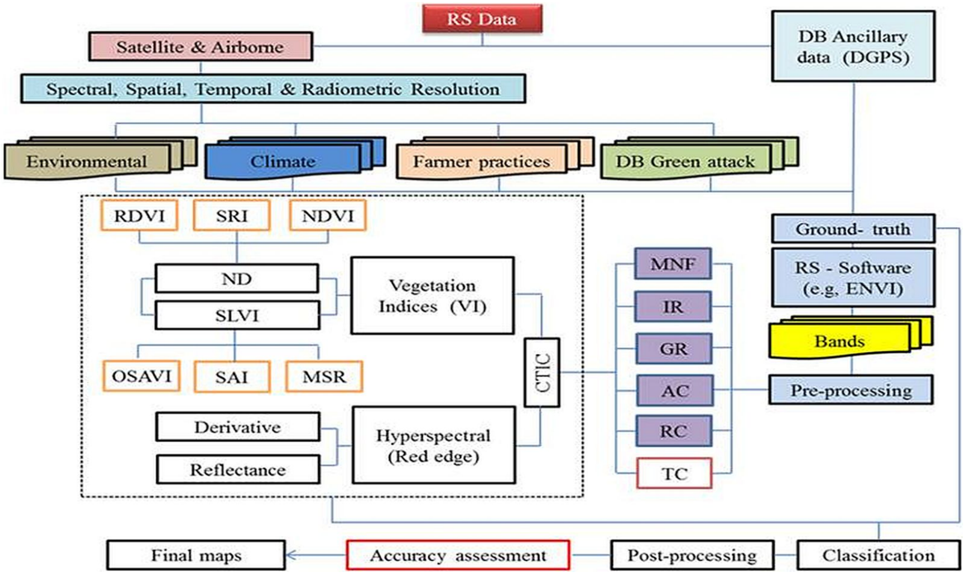


***Figure:***

*(Literature Review)*

***Few major Vegetation Indices are:***

- Simple Ratio
- Normalised Difference Vegetation Index (NDVI)
- Pigment Specific Simple Ratio(PSSR)
- Chlorophyll Index
- Perpendicular Vegetation Index
- Greenness Above Bare Soil
- Moisture Stress Index

My first  task was to develop codes for the vegetation indices.The task was accomplished with the use of Python as a programming language.Codes for around 30 different vegetation indices were developed and used for further analysis. Apple fruit, Withania Somnifera(Jambhul) and Jawar Leaf were the species on which the indices were developed.

The database used for this was mostly of ASD format which was then extracted into a structured format by exporting it through the 'ViewSpecPro' software.Further the dat/txt data was converted to csv file and imported for further use and development of indices.

- ***Derivative-based Indices**:*

Derivative-based indices analyze the slope and curvature of reflectance curves rather than reflectance values themselves and have been most useful for analysis of the "red edge" in the spectral response of vegetation (Datt and Paterson, 2000). Codes were developed for around 10 different derivative based indices and were used for further analysis.

***Table**:Major Vegetation Indices and their formulae*

| Sr. No. | Vegetation Indices | Formulae | Indicator | References |
|---|---|---|---|---|
| 1 | Normalized Difference Index | mSR = (R750-R705)/(R750+R705 ) | Chlorophyll Content | Gitelson & Merzlyak (1994) |

| | | | | |
|---|---|---|---|---|
| 2 | Modified Normalized Difference Index | mND = R750-R705/R750+R705-2*R445 | Chlorophyll Content | Sims and Gamon (2002) |
| 3 | Structure Insensitive Pigment Index | SIPI= (R800-680) / (R800+R680) | Carotenoid: Chlorophyll a ratio | Peñuelas et al. (1995) |
| 5 | Normalized difference vegetation index | NDVI = (R800 - R670) / (R800 + R670) | Biomass, leaf area | Rouse et al. (1974) |
| 7 | Photochemical Reflectance Index | PRI=R531-R570/R531+R570 | Xanthophyll Response to light photosynthetic efficiency. | Gamon et al. (1992) |
| 8 | Pigment Specific Simple Ratio | PSSRa = R800/R680 | Chlorophyll a | Blackburn (1998a) |
| 9 | | PSSRb = R800/R635 | Chlorophyll b | Blackburn (1998a) |
| 10 | | PSSRc = R800/R470 | Carotenoid | Blackburn (1998a) |
| 11 | Modified chlorophyll absorption reflectance index | MCARI = [(R700 - R670) – 0.2*(R700 – R550)]*(R700/R670) | Chlorophyll content | Daughtry et al. (2000) |
| 12 | Simple Ratio Pigment Index | SRPI=R430/R680 | Carotenoid/chlorophyll a content | |
| **Water Indices** | | | | |
| 13 | Water Index | WI = R900/R970 | Water Content | Peñuelas et |

## CORRELATION MATRICES AND CURVE-FITTING

After successful implementation of codes for the vegetation indices, the next task was to find correlation within the indices and make a proper curve-fit for the same. During this analysis, Normalized Difference Vegetation Indices(NDVI) was taken as a base index.

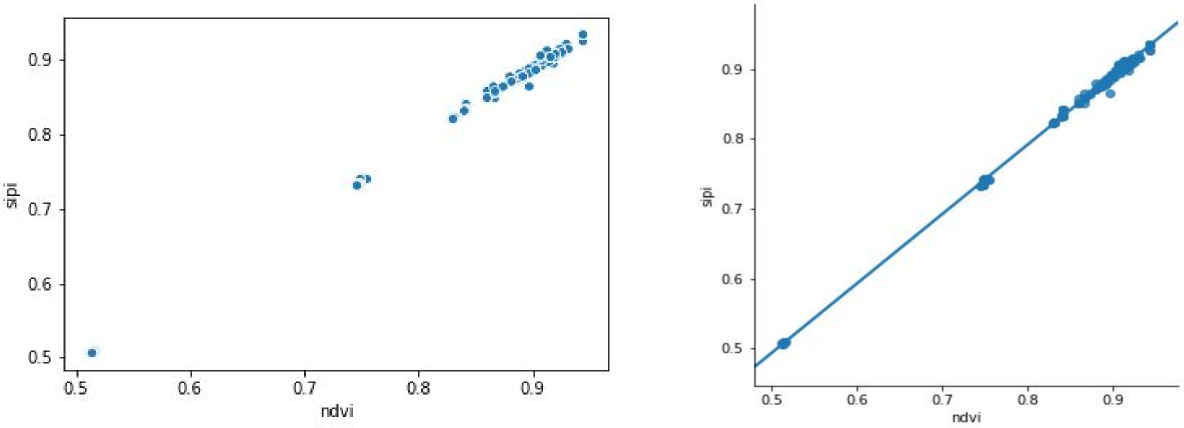
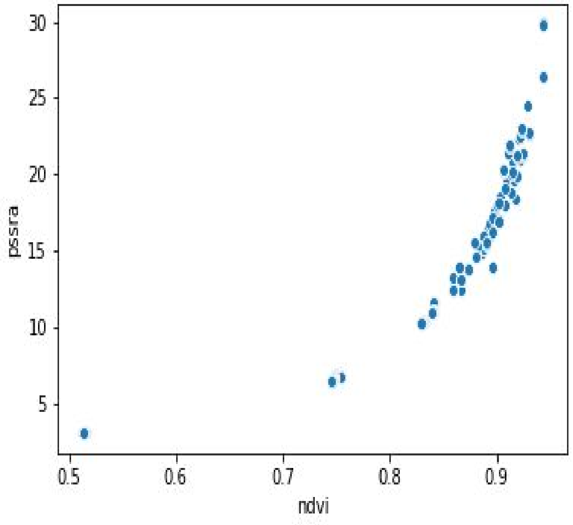
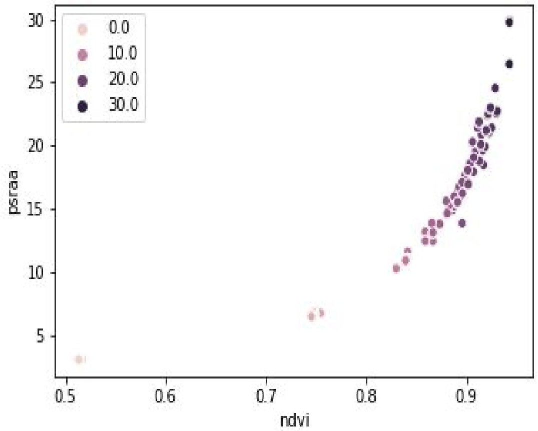
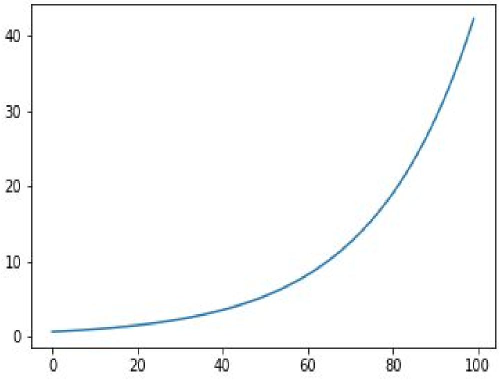


***Fig:*** *Scatterplot and corresponding curve-fit for SIPI vs NDVI plot(R2_score=0.996)*

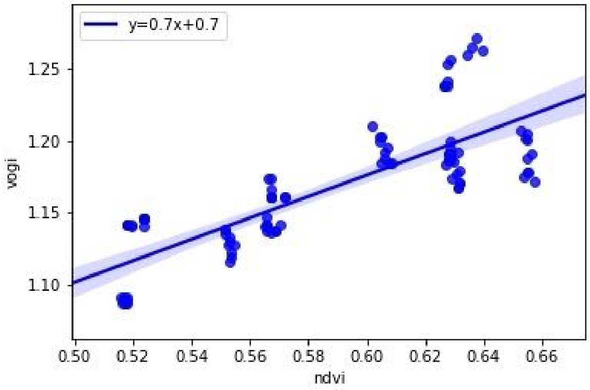


*Scatter-plot and corresponding curve fit for PSSRa vs NDVI plot. The R2_score for curve fit is 0.985*

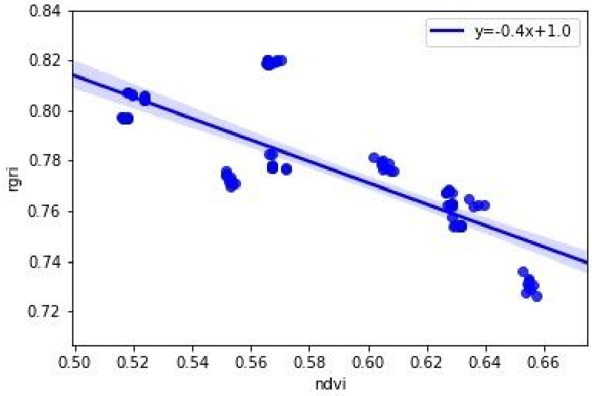




: **PSSRa= 0.0105e^(8.2944*NDVI)**



*Scatter-plot and corresponding curve fit for VOGI vs NDVI plot. The R2_score for curve fit is 0.785*



*Scatter-plot and corresponding curve fit for RGRI vs NDVI plot. The R2_score for curve fit is 0.76*

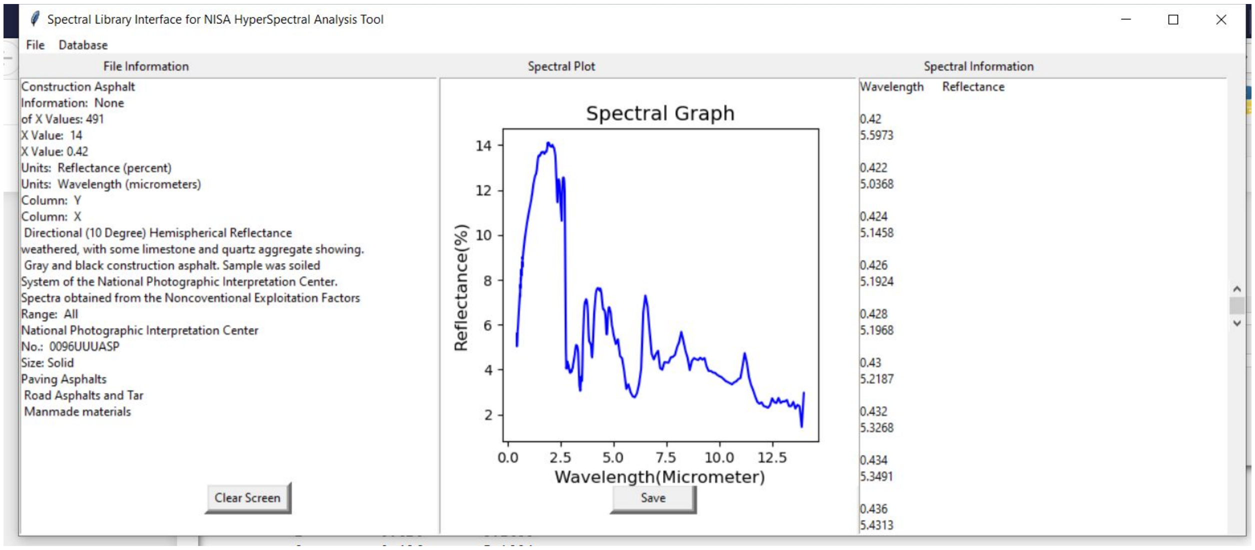
Above curves are a few instances of the curve-fitting. Similarly, scatter plotting and curve-fitting was carried out for all the indices with NDVI as base index and R2_score was determined. It is observed that there's a very high correlation with NDVI and Structure Insensitive Pigment Index(SIPI, R2=0.99) , PSSRa( R2=0.985) ,MRESR (R2=0.71) , VOGI (R2=0.68).

It is also observed that PSSRa i.e a robust index for chlorophyll determination exhibits an exponential relation with NDVI, while most of the other indices give a good linear-fit.

## DEVELOPMENT OF GRAPHICAL USER INTERFACE(GUI)

An attempt to develop a Spectral Library Interface(GUI) representing the brief information of the vegetation or any other species, corresponding spectral plot and the reflectance corresponding to wavelength was done. The GUI was designed so that it is compatible with ASD, ASTER, ASCII, NISA files.

Tkinter which is a standard GUI library for Python was used for the purpose.Python when combined with Tkinter provides a fast and easy way to create GUI applications.  Task of completing the same work using PyQt as a library is in progress and will be accomplished soon.



```
Construction Asphalt
Information:  None
of X Values: 491
X Value:  14
X Value: 0.42
Units:  Reflectance (percent)
Units:  Wavelength (micrometers)
Column:  Y
Column:  X
 Directional (10 Degree) Hemispherical Reflectance
weathered, with some limestone and quartz aggregate showing.
 Gray and black construction asphalt. Sample was soiled
System of the National Photographic Interpretation Center.
Spectra obtained from the Noncoventional Exploitation Factors
Range:  All
National Photographic Interpretation Center
No.:  0096UUUASP
Size: Solid
Paving Asphalts
 Road Asphalts and Tar
 Manmade materials
```

```
Wavelength     Reflectance

0.42
5.5973

0.422
5.0368

0.424
5.1458

0.426
5.1924

0.428
5.1968

0.43
5.2187

0.432
5.3268

0.434
5.3491

0.436
5.4313
```

## MACHINE LEARNING APPROACH FOR RETRIEVAL OF LEAF CHLOROPHYLL CONTENT

Chlorophyll present in green leaves, is a key driver of photosynthesis through its ability to convert sunlight into the biochemical energy responsible for carbon fixation.

Chlorophyll works as an indirect measure of the gross primary productivity of an ecosystem due to its robust relationship with vital biophysical and biochemical processes. The accurate estimation of leaf chlorophyll content (Chlt) is an important element in monitoring overall plant health, managing fertilizer application, as well as other inputs in agricultural systems, where productivity levels are directly related to plant condition.

The **objective of the present study** was to evaluate the potential of hyperspectral data to quantify Chlt in jawarleaf, using traditional statistical approaches in conjunction with machine learning techniques.

We benchmark the **Random Forest and XGBoost** machine learning techniques performance relative to that of a simple linear regression against individual indices.

- ***Chlorophyll Determination:***

Leaf Chlt was determined by collecting samples from the point leaves (and corresponding to the spectral sampling) via chemical extraction and spectrophotometric analysis in the laboratory.Pigment contents were determined using the methods of Arnon and Wellburn.

- ***Extraction of Vegetation Indices:***

Around 20 different vegetation indices were extracted and analysed.Out of the 20 indices used for analysis 6 were derivative indices.All of the data processing and calculations of the VIs were performed in Jupyter Notebook with Python as a programming language.

## APPROACH

At first, **simple linear regression of the Chlt against the 20 unique vegetation indices** calculated from the hyperspectral data was employed.

Following this, **an implementation of a random forest machine learning approach** was used to examine:  the use of established vegetation indices as input variables to infer Chlt, initially using the 20 selected indices and then progressively reducing the number of input variables to observe the impact on retrieval accuracy. The Jupyter Notebook software platform was used for both the data analysis and the implementation of the machine learning algorithm.

- ***Simple Univariate Regression Analysis***

The indices were evaluated for their performance in estimating leaf Chlt by undertaking simple regression analysis and curve fitting on the data obtained from laboratory analysis of leaf samples and the in situ collected spectral data. During the analysis, each of the VIs were used as a single explanatory variable one-by-one for estimation of Chlt.

Several regression models were examined, and the best-performing models were chosen based on the coefficient of determination (R2_score). Due to lack of sufficient number of samples for total chlorophyll content a good fit couldn’t be made while implementing this approach.

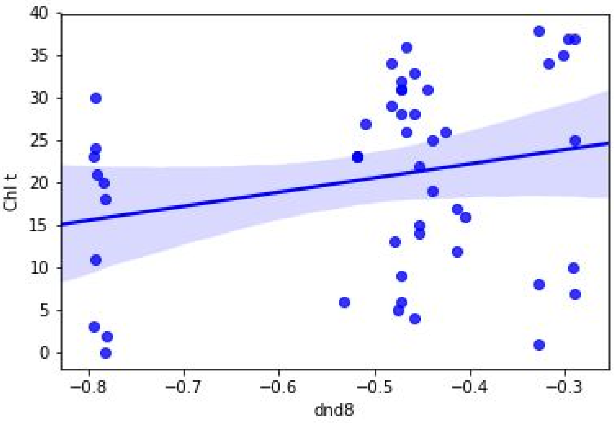


***Fig**: Linear fit model for Chlt vs dnd8 index(derivative index) (R2_score:0.12)*

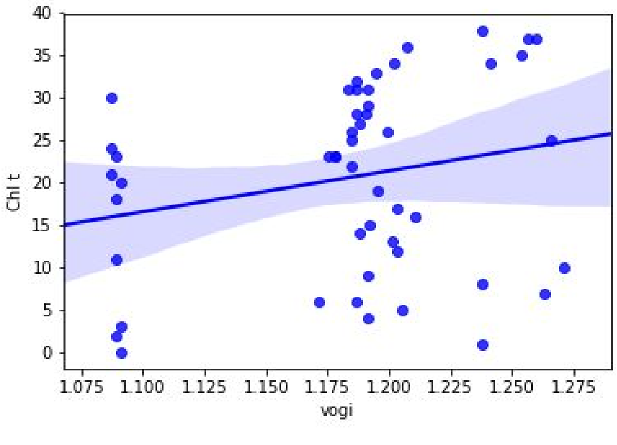


***Fig**: Linear fit model for Chlt vindex (R2_score=0.2)*

- ***Description of the Random Forest Approach***

The RF machine learning method provides a straightforward approach of feature selection and of cascading the variable importance. There are relatively few assumptions attached to RF, so data preparation and model parameterization is less challenging.

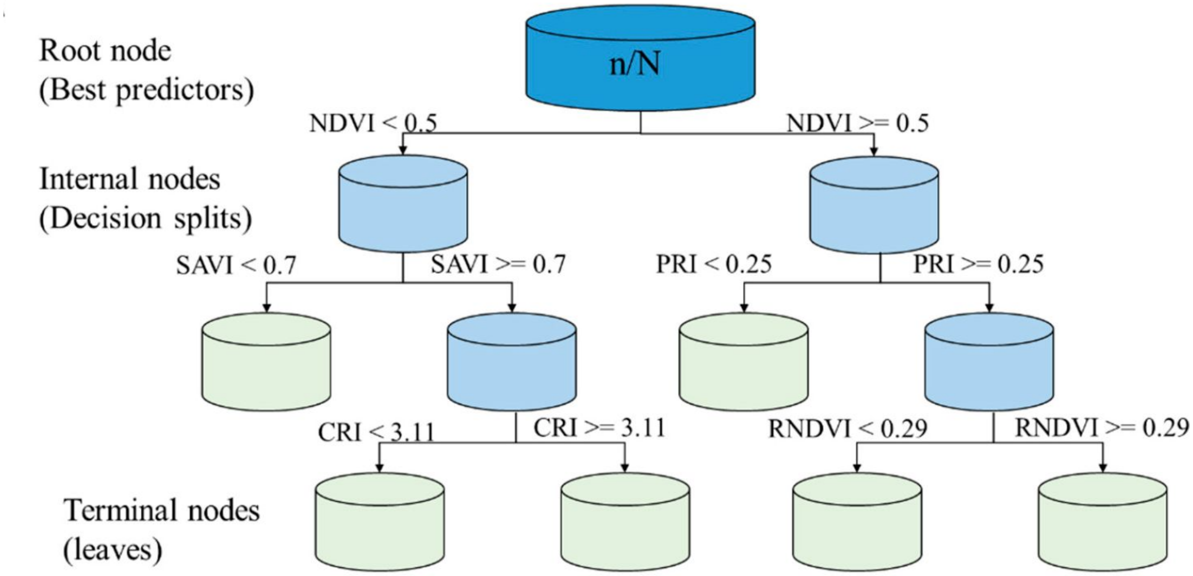


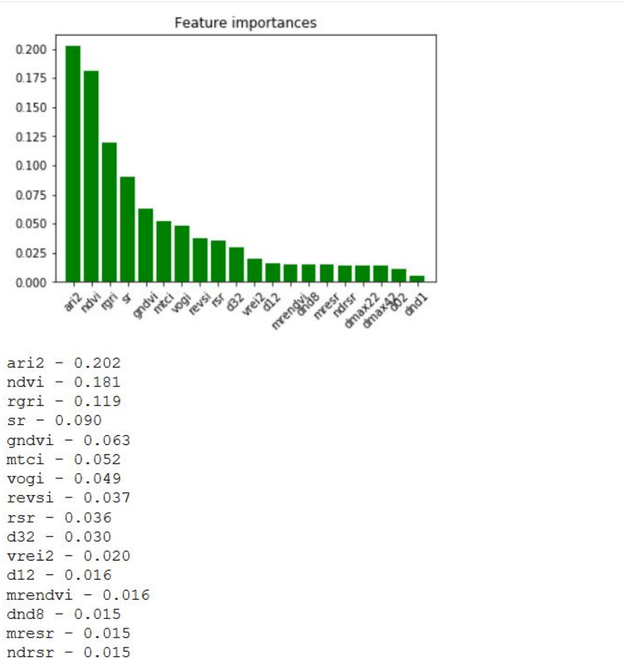
**Figure**:. Simple illustration of decision trees regression models, showing the building blocks for the Random Forest . Random Forest combines multiple randomized decision trees into a single output.

The **relative importance of the input features** was measured at the completion of the RF model training. The variable importance is based on the concept that if the exclusion of a variable is associated with a considerable reduction in prediction accuracy, then that variable is deemed as important. Thus,  a subset of attributes is selected based on important scores.

# RESULTS

Once the RF model was optimized using the selected input features (i.e., the 20 vegetation indices), they were ranked based on their importance using a forward selection function to identify which vegetation index predicts Chlt with the greatest accuracy.

The model was initially applied to the training data containing all 20 vegetation indices, based on out-of-bag permuted predictor estimates. Results show that the most important predictor was the Anthocyanin Reflectance Index (ARI2), followed by the Normalized Difference Vegetation Index (NDVI), Red Green Reflectance Index(RGRI) , the Simple Ratio(SR).**The results are found to be a good match with a recent research paper** (published in 2019, attached in References)



```
ari2 - 0.202
ndvi - 0.181
rgri - 0.119
sr - 0.090
gndvi - 0.063
mtci - 0.052
vogi - 0.049
revsi - 0.037
rsr - 0.036
d32 - 0.030
vrei2 - 0.020
d12 - 0.016
mrendvi - 0.016
dnd8 - 0.015
mresr - 0.015
ndrsr - 0.015
```

***Feature-Importance:***

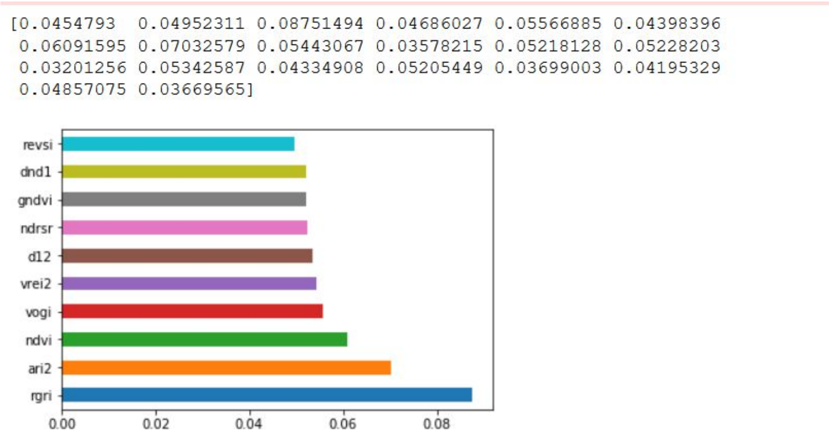
ARI2-0.202

NDVI-0.181

RGRI-0.119

SR-0.09

**Figure**: Feature Importance Ranking using RFRegressor Algorithm



```
[0.0454793  0.04952311 0.08751494 0.04686027 0.05566885 0.04398396
 0.06091595 0.07032579 0.05443067 0.03578215 0.05218128 0.05228203
 0.03201256 0.05342587 0.04334908 0.05205449 0.03699003 0.04195329
 0.04857075 0.03669565]
```

**Figure**: Feature Importance Classification using ExtraTreesClassifier

Accuracy for training the model over different depth of trees by varying the max_depth parameter and n_tree parameter was **sufficiently high ( ~0.78)** But,because of insufficient data availability for Total Chlorophyll content, even after good training and learning the model couldn’t be tested and verified on large amount of data.Attempts to gather sufficient data for chlorophyll a, chlorophyll b and total chlorophyll are in progress and later the model will be evaluated on the gathered data.

## SPECIAL THANKS

1. Mr. AmarSinh Varpe Sir
2. Mr. Dhananjay Nalawde Sir

## REFERENCES

- *John Nay, Emily Burchfield & Jonathan Gilligan (2018) A machine-learning approach to forecasting remotely sensed vegetation health, International Journal of Remote Sensing, 39:6, 1800-1816, DOI: 10.1080/01431161.2017.1410296*
- *International Journal of Remote Sensing, 2015 Vol. 36, No. 12, 3114–3133, http://dx.doi.org/10.1080/01431161.2015.1054959*
- *Wang L, Chang Q, Yang J, Zhang X, Li F (2018) Estimation of paddy rice leaf area index using machine learning methods based on hyperspectral data from multi-year experiments. PLoS ON 13(12)e0207624.https://doi.org/10.1371/journal.pone.0207624*
- *A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat (Syed Haleem Shah 1,* , Yoseline Angel 1 , Rasmus Houborg 2 , Shawkat Ali 3 and Matthew F. McCabe 1)*
- *Spectral signatures of sugar beet leaves for the detection and differentiation of diseases.August 2010, Volume 11, Issue 4, pp 413–431.A.-K. Mahlein,U. Steiner,H.-W. Dehne,E.-C. Oerke.*