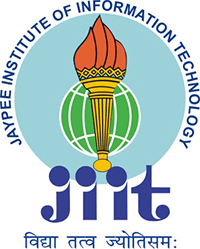
**MINOR PROJECT**



Submitted By :

Anirban Dutta 16103111 B1 : [anirbanduta98@gmail.com](mailto:anirbanduta98@gmail.com) 9711664398

Anushka Mittal 16103132 B1 : [mittalanushka5@gmail.com](mailto:mittalanushka5@gmail.com) 7011671201

Shishir Khandelwal 16103236 B1 : [shishir29121997@gmail.com](mailto:shishir29121997@gmail.com) 9810648462

Adarsh Srivastava 16103017 B7 : [daujiadarsh123@gmail.com](mailto:daujiadarsh123@gmail.com) 9454276839

**DROUGHT PREDICTION AND LAND COVER CLASSIFICATION IN AGRICULTURE USING MACHINE LEARNING**

**1. ABSTRACT**

Due to ever increasing running costs of Agriculture, farming households are struggling to make ends meet. This is a challenge for the development of the country as agriculture is highly critical for the development of the country. A lack of cost effective innovations and better tools and techniques have added to the woes of the farming households. The project aims to use Machine Learning tools to help the farmer in making a better decision about what kind of land should be used for growing a particular crop at a particular time. Detailed analysis is carried out on relevant data such as weather conditions, area and agricultural inputs for particular locations. Based on this data, project targets at the major issue of droughts by correctly predicting it’s possibility in an area. A novel approach of classifying land cover from series of multi-spectral satellite images is used which helps the farmer in identifying the correct type of land area. Such analysis and predictions can lead to overall higher productivity for farmers.

**2**. **INTRODUCTION**

Proper collection of relevant raw data, using the data to carry out predictions and analysis will provide insights about the ongoing trends and suggest proper measures. The projects aims to extract the raw data of drought affected areas of certain districts of Maharashtra and further empower the farmers to become aware about the future climatic changes. Extracting accurate data and implementing correct algorithms can help formulate solutions to lessen the effects of droughts in drought prone states. After prediction of droughts, we would move further to the second phase, i.e. Land Cover Classification. Land Cover Classification is done on the basis of satellite images taken from sites like Kaggle, UCI Machine Learning Repository. Before feeding into any kind of machine learning model, these images are preprocessed so as to obtain better results. Each image fed into the classifier is labelled into its corresponding class. Certain ways have been implemented to study the satellite images and the tags corresponding to them. Cluster analysis, t-SNE Embedding and NDVI index help to analyse the images.

**Contribution to Sustainable development:**

Factors such as concern for food security, ever increasing population and climate change has made use of innovative approaches such as AI, Machine Learning a necessary to improve and protect the crop yield. Such approaches can help farmers learn about the techniques they should employ. Detailed predictions and Analysis can be carried out to help the farmers in making better decisions.

**3. RELATED WORK**

Various methods have been implemented to predict droughts using stochastic linear models, artificial neural networks. Certain extraction learning models were applied on satellite images as well. Some of them are discussed below.

# Authors have implemented the prediction of droughts and climate changes for future years. They have made use of dataset of satellite images for 24 years on vital parameters. 10 different models were implemented to predict the probability of drought after every 4 months. [2]

# 

# The concerned authors have discussed about forecasting future droughts in Eastern Part of West Bengal. They made use of stochastic linear models like ARIMA and SARIMA. The results were compared with real time data. The models work accurately for predicting the probability of droughts for the next 1-2 months ahead. These models have shown potential to forecast reliably towards the goal of drought forecasting. [3]

# 

# The paper discusses a different approach to predict drought using Artificial Neural Networks and Matlab Neural Toolbox. Data was extracted for 4 weather stations of 30 years and different combinations of input output were applied on the model to achieve effective outcomes. [5]

# 

# Author applied Convolution Neural Network models for classifying the land cover area into the corresponding labels. Models were implemented on UC Merced Dataset. Moreover, they also introduced an efficient algorithm to handle large dimensional images. [4]

# 

# Authors make use of deep learning models for classification of satellite images. The data was applied on CaffeNet and GoogLeNet with varying learning parameters. Techniques were applied on data to remove overfitting and reduce compile time. The models showed significant improvement and applicability. [8]

# 

# Authors make use of the deep learning models combining CNNs with extraction learning models (ELM). The model was tested on UC Merced dataset consisting of 21,000 images belonging to 21 classes. Models applied gave accurate and satisfactory results.

**4. DESIGN**

**4.1 Dataset**

**1) Dataset for the drought prediction phase**

The features of the dataset for this phase have been scrapped from various governmental sites. Pressure data has been scraped from the site timeanddate [7] using BeautifulSoup from 2001-2018. While the rainfall data has been taken from the maharain governmental site [12]. Temperature conditions have also been collected from thetimeanddate site [13]. These weather conditions are extracted for regions shown in Figure 1 are then integrated for predicting the drought labels.

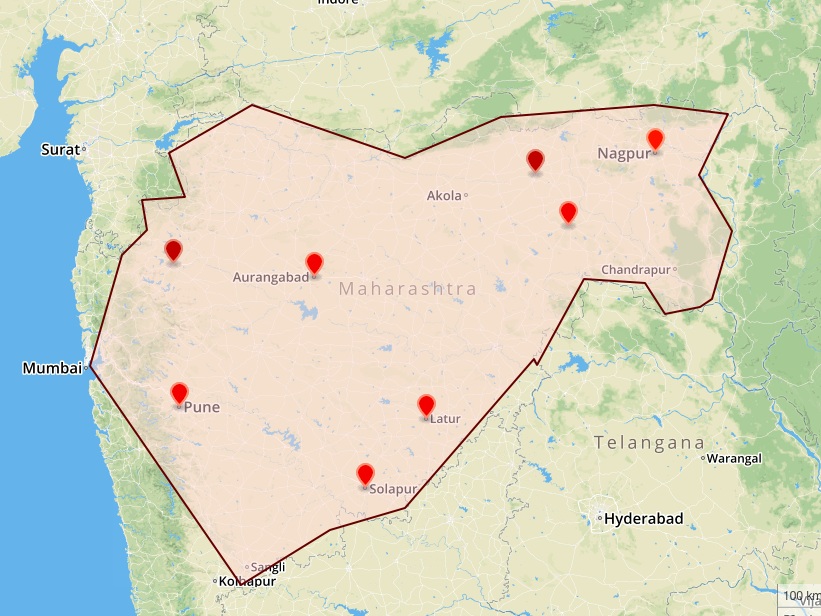
.

Figure 1 : Drought affected districts of Maharashtra

**4.2 Module Description**

The dataset of drought prediction is highly imbalanced. Several oversampling and undersampling techniques have been implemented to remove the imbalance and classify the drought labels correctly.

**5) IMPLEMENTATION**

After successfully scraping data, binning was applied on the raw data using python tools which helps to convert the continuous data to categorical. The method was implemented on all parameters like pressure, temperature and rainfall. After binning, the drought labels were calculated and analysed for the chosen districts from 2001-2018. For feeding the dataset into a classifier, all the parameters were combined together and mean rainfall, temperature and pressure was calculated as well. After analysing the final dataset, it was observed that the dataset was imbalanced. To remove the imbalance undersampling and oversampling was implemented to make the dataset balanced using SMOTE python tool.

**5.1 Methodology**

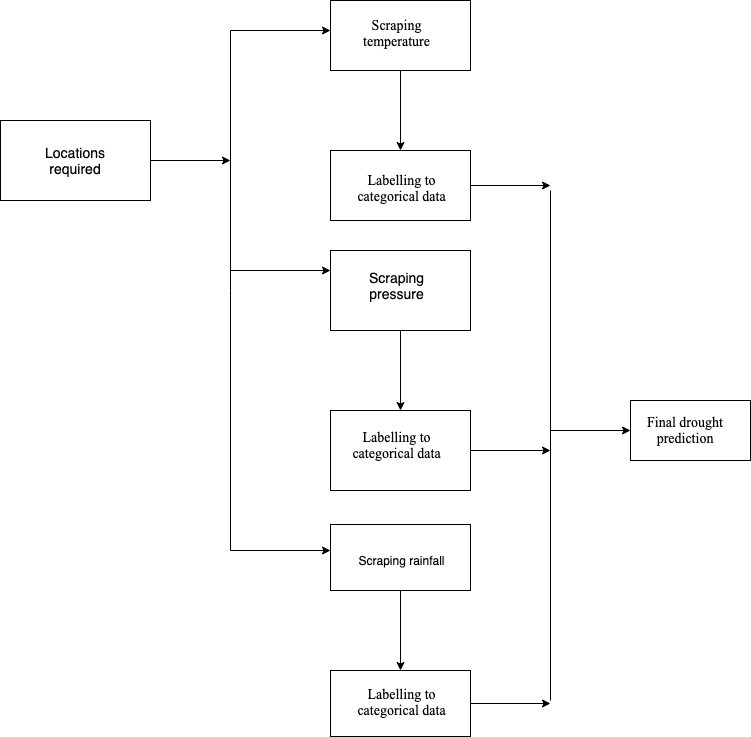
****

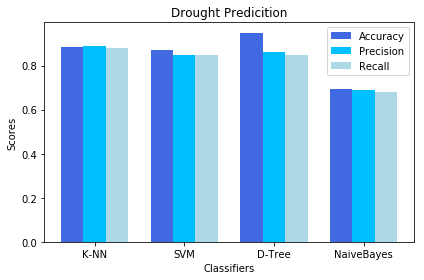
Figure 2 : Flowchart for the drought prediction

The workflow of the drought prediction phase is shown in the above flow chart.

For the drought prediction the dataset was fed into several classifiers like SVM, K-NN, Decision Tree, Naive Bayes.

**5.2) Results**

The drought prediction dataset was fed into classifiers and parameters like accuracy, precision, recall, specificity and sensitivity was calculated.



Graph 1 : Classification Results of drought predicted regions

The graph above depicts the comparative study between classifiers and their performances. It is observed that Decision Tree Classifier provides the maximum accuracy of 95%. Other essential performance metrics are shown below. Naive Bayes fails to classify the droughts correctly with a poor accuracy of 69%.

a) Classifier :

Decision Tree Classifierprovides an accuracy of 95% with specificity of 95 % and sensitivity of 87%

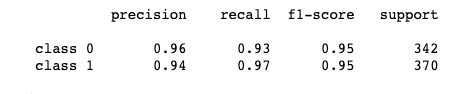
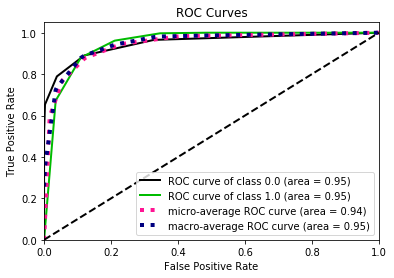
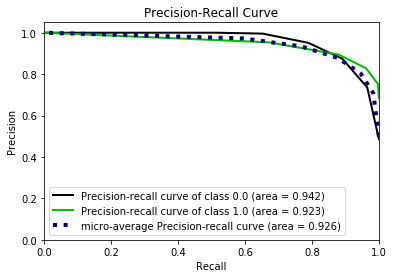


Figure 3 : Classification report for the decision tree classifier



Graph 2 :ROC Curve for decision tree classifier

The Roc Curve in graph 2 clearly depicts that the classifier detects more class 0 labels as compared to class 1 labels. The roc curve of class 0 is more towards the True Positive Rate and away from False Positive Rate. Moreover, the calculated specificity (0.95) is higher than the sensitivity (0.87) while proves our analysis to be accurate.



Graph 3 : Precision-Recall curve for decision tree classifier

The precision recall curve is used while handling binary classifications. Because of the minor imbalance between the classes, it can be inferred from the graph that area under curve of class 0 is more that area under curve of class 1. Therefore, the precision recall of class 0 is higher than that of class 1.

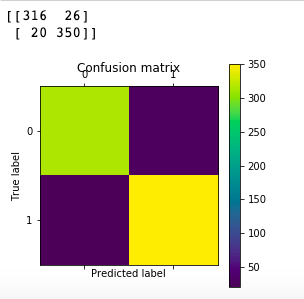


Figure 4 : Confusion matrix for decision tree classifier

The figure 4 depicts the count of True Positive (316), False Positive (26), False Negative (20), True Negative (350) evaluated by Decision Tree Classifier on the drought prediction dataset.

**PHASE 2: LAND COVER CLASSIFICATION USING SATELLITE IMAGES**

**4. DESIGN**

**4.1 Dataset : UC Merced dataset**

This is a 21 class land use image dataset whose images have been manually extracted from the USGS National Map Urban Area Imagery collection for various urban areas around the country at a pixel resolution of 1 foot.

The dataset has been downloaded from the UCI Machine Learning Repository **[**14**].**

The images from the dataset are shown below-



Figure 5 : Images in the UC Merced dataset

: **Amazon rainforest dataset**

This is a 17-class satellite based image dataset having classes like hazy, primary, agriculture, cloudy. The task of the project is to provide tags to each image in the dataset, which are segments of a larger image in the Amazon Rainforest. Each image measures 256\*256 pixels. However, it has been observed that each image can have multi labels, which makes it a multi-label classification as compared to the standard multi-classification. Some of the images with their tags are shown below-

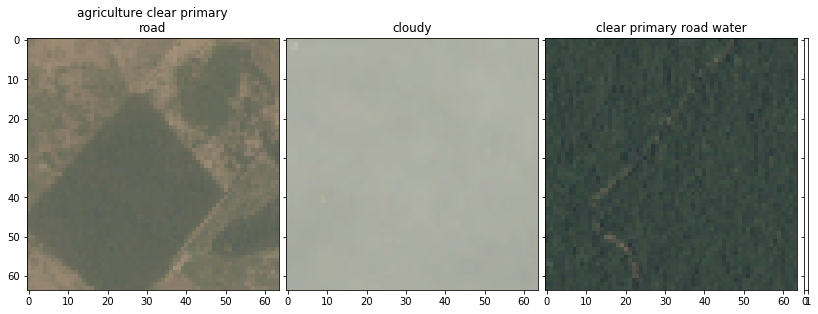
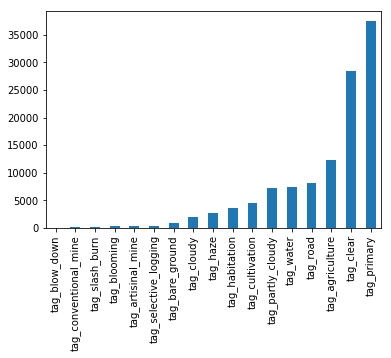


Figure 6 : Images in the Amazon rainforest dataset

**4.2 Module description**

The UC Merceddataset consists of 2100 images with 100 images belonging to each of the 21 classes. Each image measures 256\*256 pixels. The output labels in the dataset include classes like agriculture,airplane, baseball, beach, river, forest, runway.

The Amazon Rainforestdataset is a highly imbalanced dataset with the Primarytag being the most prominent tag. The distribution of the different classes in the dataset is shown in the graph below.



Graph 4 : Distribution of tags in the dataset

From graph 4 it is evident that tag\_primary has the maximum count and the tags like cloudy, tag\_blooming, slash\_burn, blow\_down are rare.

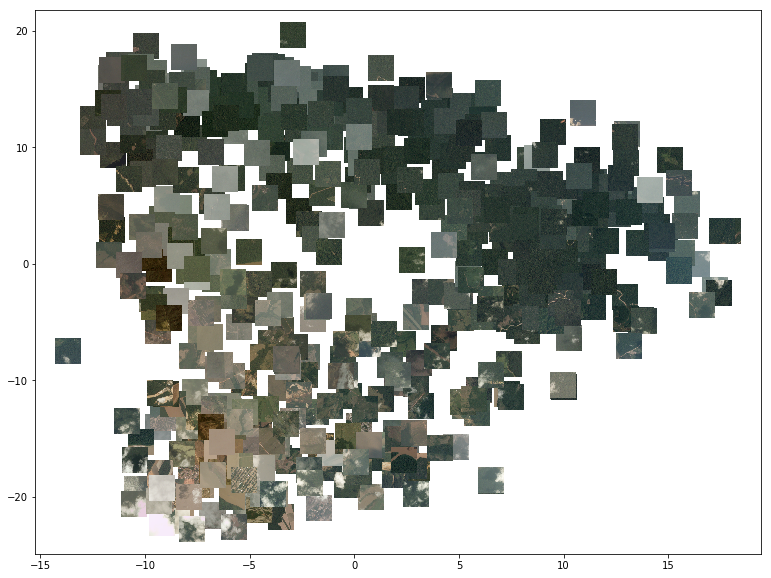


Figure 7 : Overview of the Amazon rainforest dataset

Figure 7 gives us a better overview of the dataset. t-SNE model has measured the distributions of pairwise similarities of input images. X- axis represents the ‘t-SNE one’ and Y-axis represents the ‘t-SNE two’ . Majority of the images which belong to rainforest primary tag can be observed by large number of green chunks of boxes. We also can observe the rare images of cloudy and hazy weather conditions. This has been generated using the OffsetImage and AnnotationBbox from matplot library.

Since, it is a multi-label classification problem, it is worth noting down the co-occurence of different tags in the dataset. The figure below shows the co-occurence matrix which depicts what percentage of the X label also has the Y label.

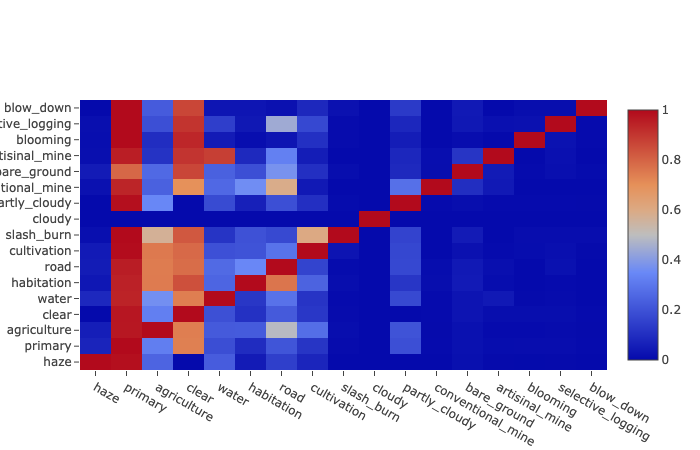


Figure 8 : Co-occurence matrix for the Amazon Rainforest dataset

**5. IMPLEMENTATION**

After successfully obtaining the images from the UC Merced dataset**,** these are preprocessed, cropped and resized. Various CNN models like VGG16, VGG19, Xception have also been applied to the images showing considerable accuracies in the final classification.

For the Amazon Rainforest dataset the images in the dataset possibly belong to one or more classes i.e. an image can have multiple tags. For example, an image train\_0.jpg in the dataset belongs to multiple tags like haze and primary and similarly the other images. The tag predictions expected for a test image are also multiple.

Every image in the dataset is assigned a list of tags it belongs to, but for the output labels to be fed into any classifier, it needs to be of standard shape. Since the count of total possible classes in the dataset is 17, every image is assigned a boolean array of size (17,) where only the values for the image tags are set to 1 and the rest to 0. So for the image train\_0.jpg**,** values corresponding to the classes haze and primary are 1 and others are 0 in the array. This is achieved using the MultiLabelBinarizer() class provided by the sklearn library.

Further after preparing the input and the output labels, these images are preprocessed ( resized and normalized) for better classification results. Different ensemble learning techniques like xgboost and deep learning models like VGG16, VGG19 have been applied on these final images and the resultant classification results are noted down. The performance of these algorithms is evaluated in terms of the metrics like accuracy, precision and recall.

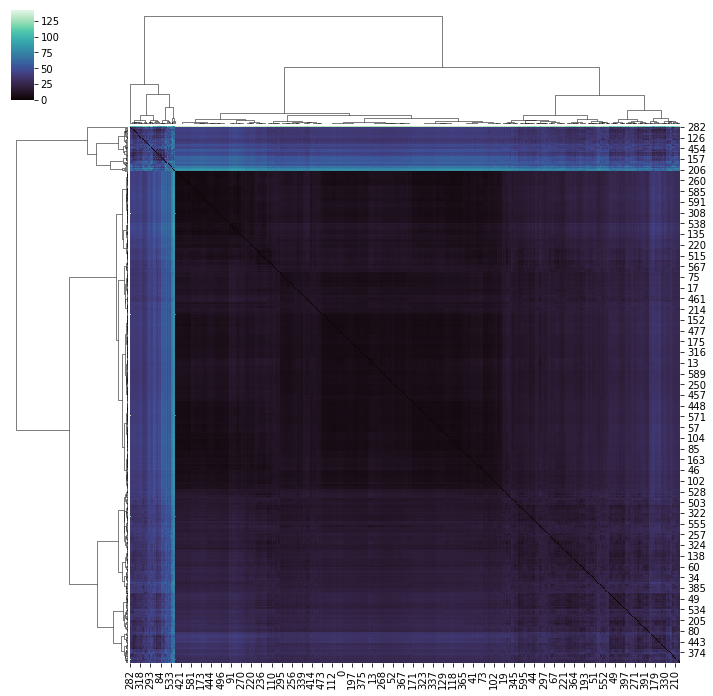


Figure 9 : Cluster Analysis of images in the Amazon rainforest dataset

The images have been clustered by their pixel intensities and computing distances pairwise. The images have been reshaped, normalised and fitted into a square form. Spatial distance library has been used for the cluster analysis. Figure 8 portrays that very few images are dissimilar to all other images by using the pixel intensities.

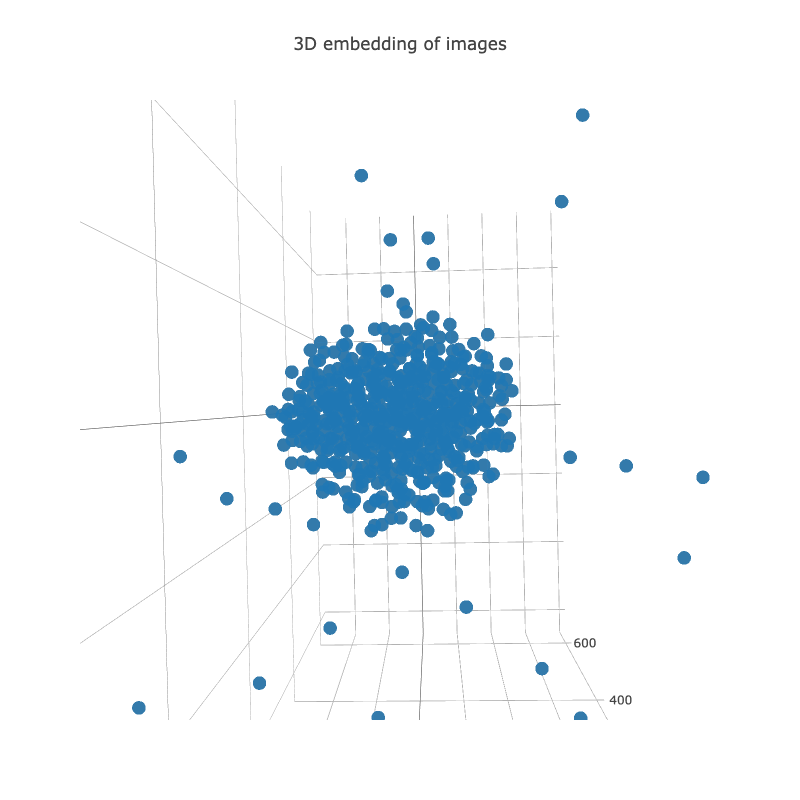


Figure 10 : 3D analysis of Amazon rainforest images

Figure 10 provides a better overview regarding the outliers and plots the images in 3D format. The t-SNE Embedding in figure 9 provides the evidence regarding the dissimilarities and the outliers. t-SNE are usually used for data exploration and visualisation.



Figure 11 : Maximum similar and dissimilar images

From the distance matrix of the images of pixel intensities, we found the most dissimilar image and the most similar image in the dataset using the nanargmax, nanmean functions and computing the average distances to all different images as shown in figure 10. As we can observe, the image with least distance cost metric is an image with primary rainforest. The image with maximum average distance depicts clouds and haziness.

**ANALYSIS (NDVI INDEX)**

The NDVI (normalized difference vegetation index) is an indicator of presence of green vegetation for a target area, it gives a measure and degree of photosynthetic capacity of vegetations.[15][16] It works on the basis of the fact that live green plants absorb solar radiations in the PAR spectral region (photosynthetically active radiation) and the leaf cells of the plants re-emit solar radiations in the near-infrared spectral regions.

The NDVI is calculated as follows:

- Equation(1)

Here,

NIR stands for spectral reflectance in red spectral region (the visible spectral region).

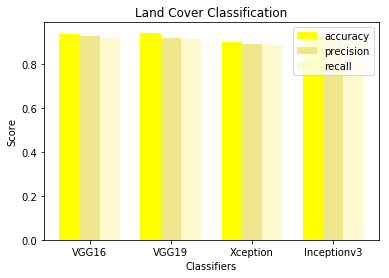
Red signifies spectral reflectance in near-infrared regions. [17]

Different areas and their usual NDVI index values:

1. Regions having a dense vegetation canopies usually have positive values in range 0.3-0.8.
2. Clouds and snow fields are indicated by negative NDVI index.
3. Water bodies show a rather low positive or slightly negative values, which can be attributed to low reflectance in both NIR and Red. E.g. of such bodies can be Oceans, Seas, Lakes.
4. Barren land or Soil covered regions generate small positive NDVI values (in range of 0.1 to 0.2).

**RESULTS**

UC Merced datasetimages were fed into several CNN models and performance metrics like accuracy, precision, recall were measured.



Graph 5 : Classification results for the UC Merced dataset images

Graph 5 depicts the comparative study between the different CNN models and their performance. It was observed that the VGG19model provides the maximum accuracy of 94.16%. Other essential metrics like precision and recall are also shown in the graph for these models. Other models like VGG16**,** InceptionV3show equally good performance with accuracies of 93.5% and 92.5% respectively.

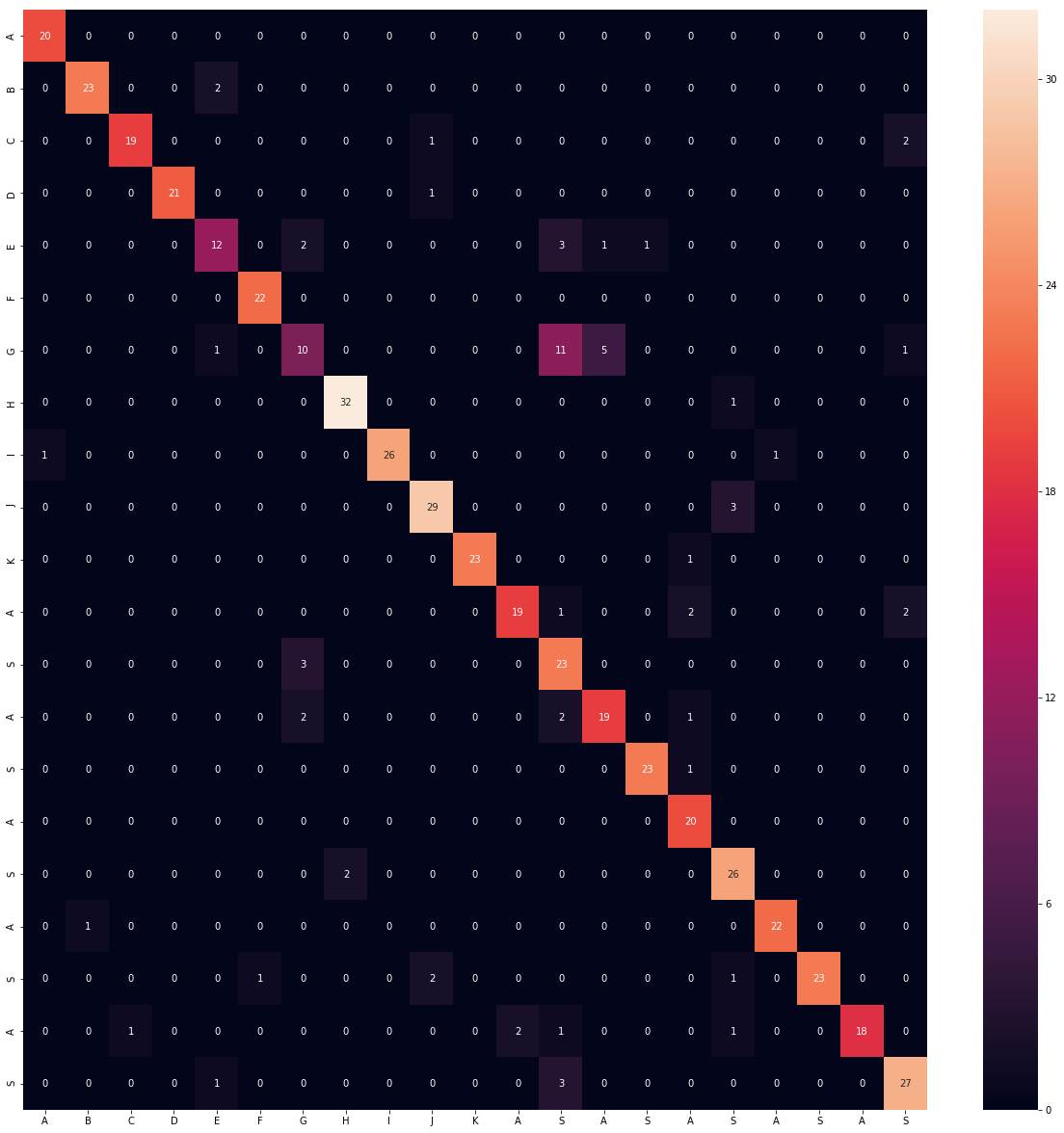
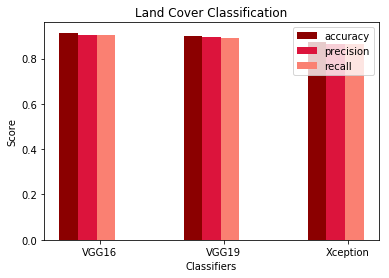


Figure 12 : Confusion matrix for VGG19 model on UC Merced dataset

Figure 12 depicts the confusion matrix as evaluated by the VGG19 model.We have the true labels along the y-axis and the predicted labels along the x-axis. The count of the correct predictions and the incorrect predictions for different classes in the dataset can be easily identified from the figure 12..

Amazon Rainforestdataset fed into the XGBoost classifier returns an accuracy of 88.9%. The predicted output for the test images are also muli-labelled. Results of models like VGG16, VGG19 and Xception on the dataset are shown in the below graph.



Graph 8. Classification results for the Amazon rainforest images

Graph 8 depicts the results obtained on applying the different models on the Amazon Rainforest dataset images. It is observed that the VGG16 model gives the best results with an accuracy of 91.4%, while the other models VGG19 and Xception also show comparable performances with accuracies of 89.9% and 87.11% respectively.

**NDVI INDEX RESULTS**

The certain satellite images very fed into the NDVI model for better analysis of vegetation of that particular area. This parameter also helps to determine the vegetation as and when the climate conditions change throughout the year. The NDVI indexes are depicted from figure (12 - 14).

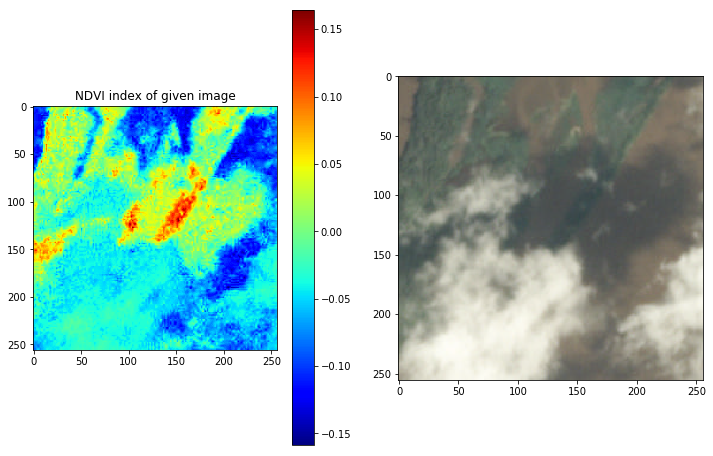
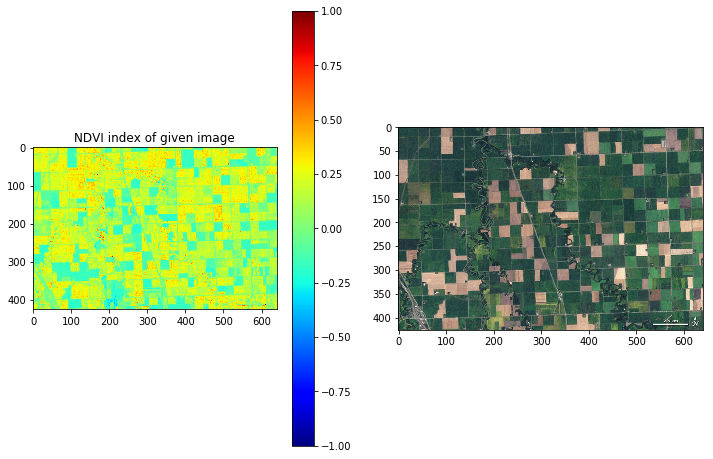
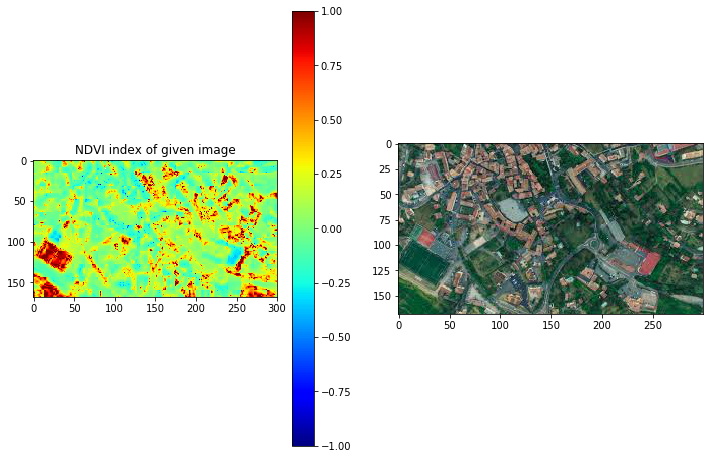
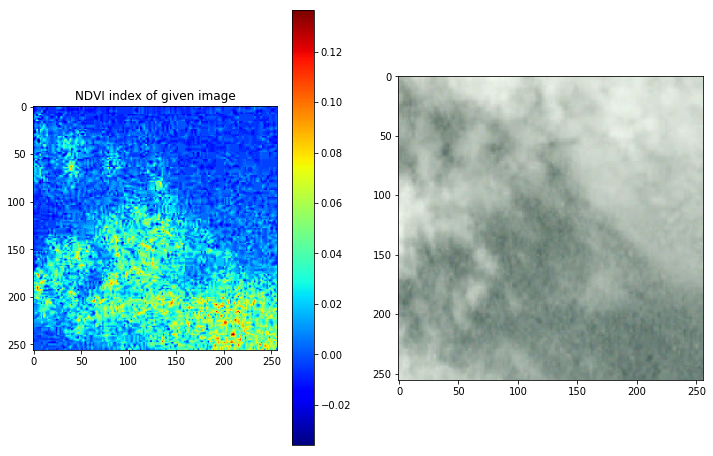
Figure 13 Figure 14

Figure 15 Figure 16

The figure 12 displays farmlands and are therefore assigned NDVI value in range 0 to 0.25, in contrast, figure 15 displaying and image containing majority of pixels showing clouds has NDVI value in range -0.02 to 0. The model performs a good job in figure 14 where the farmlands having green vegetation have been correctly assigned NDVI in range of 0.5 to 1.0 while the urban buildings have been assigned NDVI in range of -0.25 to 0. Hence, the model is able to fairly differentiate between the present regions. However, in figure 13, due to the clouds and the shadows they cast on land, some areas have been misclassified, the programme marks the cloud filled areas with values of -0.15 to -0.05.

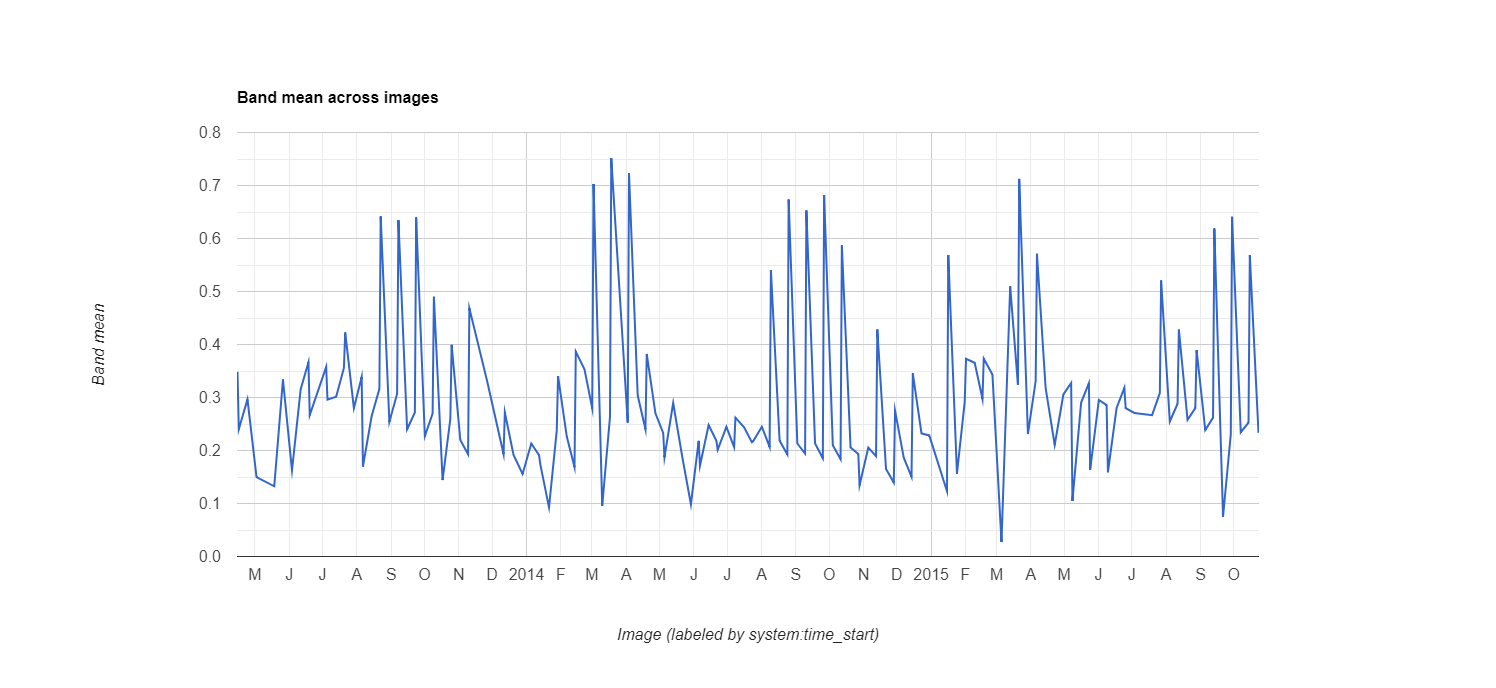
Figure 17 : Year wise NDVI Analysis 

Figure 17 is generated using Google Earth Engine. The NDVI index trend for a selected area has been depicted from 2013 to 2016. The trend also gives insights of the months of a year when the vegetation is more compared to other months.

**6. Conclusion**

The projects works in two phases with first phase being the drought prediction for the districts of Maharashtra using various Machine Learning classifiers. It has been observed that among several models applied, Decision Tree gives the best accuracy on testing. Moreover, the technique used for oversampling of data has put a larger impact on accuracy. Removing the imbalance is vital as it may lead to overfitting of data of the rare classes. The second phase employs the methods for classification of remotely sensed data or the satellite imagery. The images taken from height of 3-4 m shows the land cover area for a particular region. Deep learning models prove to be the most efficient for dataset with images and hence different CNN models are explored using which the land cover is classified. Both the phases combined serves as a great recommendation system for farmers helping them increase their productivity and hence a great tool towards sustainable development. The NDVI index acts as an useful tool to assess the vegetation trends in a region.

**References**

**1**. Aiyelokun, Oluwatobi & Ogunsanwo, Gbenga & Oluwafunbi, Fabiyi. (2018). An artificial intelligence based drought predictions in part of the tropics. Journal of Urban and Environmental Engineering. 11. 10.4090/juee.2017.v11n2.165173.

2.Berhan, G., Hill, S., Tadesse, T., & Atnafu, S. (2014). Drought prediction system for improved climate change mitigation. *IEEE Transactions on Geoscience and Remote sensing*, *52*(7), 4032-4037.

3. Mishra, A. K., & Desai, V. R. (2005). Drought forecasting using stochastic models. *Stochastic Environmental Research and Risk Assessment*, *19*(5), 326-339.

4. Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, *13*(1), 105-109.

5.Masinde, M. (2014). Artificial neural networks models for predicting effective drought index: factoring effects of rainfall variability. *Mitigation and adaptation strategies for global change*, *19*(8), 1139-1162.

6. Borra, S., Thanki, R., & Dey, N. (2019). *Satellite Image Analysis: Clustering and Classification*. Springer.

7 .North, H. C., Pairman, D., Belliss, S. E., McNeill, S. J., Cuff, J., & Hill, Z. (2013, July). Spectral classification of crop groups for land use identification with temporally sparse time-series satellite images. In *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS* (pp. 4237-4240). IEEE.

8. Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*.

9. Singh, P., & Borah, B. (2013). Indian summer monsoon rainfall prediction using artificial neural network. *Stochastic environmental research and risk assessment*, *27*(7), 1585-1599.

10. Berhan, G., Hill, S., Tadesse, T., & Atnafu, S. (2014). Drought prediction system for improved climate change mitigation. *IEEE Transactions on Geoscience and Remote sensing*, *52*(7), 4032-4037.

11. Weng, Q., Mao, Z., Lin, J., & Guo, W. (2017). Land-use classification via extreme learning classifier based on deep convolutional features. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 704-708.

12. https://www.timeanddate.com/

13. http://maharain.gov.in/

14.http://weegee.vision.ucmerced.edu/datasets/landuse.html

15.PlanetUnderstandingtheamazonfromspace/Kaggle/https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview/evaluation

16. Sellers, P. J. (1985). Canopy reflectance, photosynthesis and transpiration. *International Journal of Remote Sensing*, *6*(8), 1335-1372.

17. Myneni, R. B., Hall, F. G., Sellers, P. J., & Marshak, A. L. (1995). The interpretation of spectral vegetation indexes. *IEEE Transactions on Geoscience and Remote Sensing*, *33*(2), 481-486.Myneni, R. B., F. G. Hall, P.J. Sellers, and A.L. Marshak (1995) 'The interpretation of spectral vegetation indexes', *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 481-486.

18. https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring\_vegetation\_2.php