|  |  |
| --- | --- |
| sugarcane prosperity for farmers | Helping Sugarcane Farming Industry Prosper  Case Studies in Data Science |

Case Studies in Data Science

**Analysis of Satellite Images for Sugarcane Fields**

Affiliated to School of Science, RMIT University

**Authors: -** **Data: -** 20stOct 2019

Radhika Santosh Zawar (S3734939)

[s3734939@student.rmit.edu.au](mailto:s3675682@student.rmit.edu.au)

Kriti Kumar (s3675682)

[s3675682@student.rmit.edu.au](mailto:s3675682@student.rmit.edu.au)

Contents

[Introduction 3](#_Toc22676126)

[The Problem 3](#_Toc22676127)

[Data Understanding 4](#_Toc22676128)

[The Proposed Solution 6](#_Toc22676129)

[Methodology 8](#_Toc22676130)

[Conclusion 16](#_Toc22676131)

[Appendix A – Project Management 17](#_Toc22676132)

[References 20](#_Toc22676133)

# Introduction

Sugar is Australia's second largest agricultural export (after wheat) which totals about $2B per year. More than 95 % of sugar in Australia comes from sugarcane and is widely grown crop. More than 48,000 people are employed in this industry with over 95% of sugarcane production localized to Queensland. Our aim is to use satellite image captured data by Sentinel 2A satellite to provide insight into the production of sugarcane over time. If we can forecast how a typical crop of sugarcane will look like in coming future. Optical image time series captured using satellite is proved to be very useful for prediction of agricultural crop. There are few limitations and hardships that need to be encountered in this approach. One of them being the fact that observations are unclear because of cloudy and foggy weather conditions. We targeted this problem using small portion of area from Proserpine, Queensland clustering the fields. We aim to build an application that can manipulate and analyze satellite images from sugarcane farm and can forecast using predictive models. The application will help various stakeholders to understand soil quality, crop quality and expected harvest for sugarcane factories by predicting harvest based on historical crop cycle. This will help farmers, government, banker and various business vendors take better informed decisions based on the predicted values. Thus, Farmers will be able to improve their crop yield and get better visibility, Bankers will be able to make decision on loans based on crop value when harvested, Government will be able identify regions and people affected by weather conditions and Fertilizer vendors to Target farms to maximize the benefit to the farmer.

# The Problem

* 1. Sugarcane industry is very vast and enormous in terms of its utilisation and stakeholders. If a farmer can plan his planting and harvesting, it will aid in better productivity and yield. Farmer should have better visibility in terms of how his crop is performing as compared to other farmer’s in the area. Soil quality in terms of nitrogen and phosphorus content is needed by both farmers and fertilizers companies to better target their products so that it can benefit the farmers most.
  2. Government faces huge problem in identification of areas which are affected by extreme weather conditions and low productivity. Forecasting using satellite images we can solve these problems as we will have better insights to identify plots suitable for sugarcane.

We need to analyze sugarcane farm time series images from Sentinel 2A Copernicus Satellite. The main aim is to accurately use the images data to predict vegetation. True color images (TCI) were used as they give better insights as compared to other bands. The Sentinel-2A (S1A) synthetic aperture radar (SAR) data featuring relatively high spatial-temporal resolution provides an ideal data source for all-weather observations. The main aim is to develop a method for mapping different seasons of sugarcane. Sugarcane mapping using SAR data is a challenge because of the complexity involved in using pixels data. For example, a non-sugarcane region can be confused with sugarcane region if the masking is not appropriate.

The images are also impacted by external factors like cloud cover which reduces the visibility. Another potential issue that can be encountered is that waterlogged surface can be confused with an irrigated sugarcane field if their temporal backscatter signatures are similar. This is especially true in the case of early season mapping of sugarcane, when only part of the time series data is available.

# Data Understanding

The data was available in phases wherein the phase one had time series of tiles, along with masks to identify sugarcane farms for a small region of Proserpine in Queensland. There was a meta data json file which had information about conditions of the capture, and its location in latitude/longitude. The Phase two had data for greater region of Proserpine with 65 tiles in 64610 image files. The tile mask generation was done using the Geo JSON code. The last phase is expected to have increased size of target area to all of Australia (or Queensland / Northern New South Wales) which is yet to be released.

Images are obtained from the TCI band which represents a colorized version of the images. Data about each capture is analyzed using the metadata file available which has information about cloud cover, snow cover, latitude/longitude and helped getting deeper insights.

After understanding the problem statement, we moved further to data preprocessing and exploration to develop deeper understanding before coming up with a solution. For data understanding we converted JSON data to excel from the metadata. The metadata gave us useful information for the capture like cloud cover percentage, snow cover, cloud cover, date of image, season which are very useful for our analysis. The process can be highlighted as below.

* Image data pixels reading and conversion into useful percentages
* Converting JSON data to Excel
* Masking Sugarcane region
* Handling cloud cover data
* Trend analysis of harvest in data
* NDVI calculation and analysis

The mask provided was applied on the TCI images to help identify areas of sugarcane compared to other crops.

Harvested Region Sample output after masking the region and applying threshold for harvested region

A close up of a flower

Description automatically generated

Figure 1 Harvested Region Sample output after masking the region and applying threshold

The JSON file was converted into one table to capture information about each capture, all the images where cloud cover was more than 50 Percent were ignored as they can impact our model created.

# The Proposed Solution

NDVI index is normalized difference vegetation index is a quantitative index that gives greenness of a field. The normalized difference vegetation index is used as a base for our analysis. It depicts greenness of field ranging from -1 to +1. NDVI in Australia ranges from 0.1 to 0.7 where higher value associated with greater vegetation and greenness. As a part of photosynthesis, the green vegetation absorbs the visible light (solar radiation) and at the same time plants reflect solar energy in the near infrared. The difference in absorption is unique to live vegetation and gives a measure of green vegetation.

NDVI is calculated using red and near-infrared reflectance rRed and rNIR

NDVI = (rNIR - rRed) / (rNIR + rRed)

The value decreases as leaves come under water stress, die or get diseased.

Using NDVI value we built a predictive model that will be good indicator of sugarcane life cycle.

We will then segregate life cycle of sugarcane in four stages using this time series model to predict future stages in the outcome. These phases are Establishment, Vegetative, Grand Growth and ripening. Theoretical NDVI Index ranges for sugarcane phase classification are as below.

* Harvested Phase: NDVI value between 0.0 to 0.25
* Establishment Phase: NDVI value between 0.25 to 0.35
* Vegetative Phase: NDVI value between 0.35 to 0.45
* Grand growth Phase: NDVI value between 0.45 to 0.55
* Ripening Phase: NDVI value above 0.55

We built an application that can manipulate and analyze satellite images from sugarcane farm and can forecast using predictive models.

The application will help various stakeholders to understand soil quality, crop quality and expected harvest for sugarcane factories by predicting harvest based on historical crop cycle.

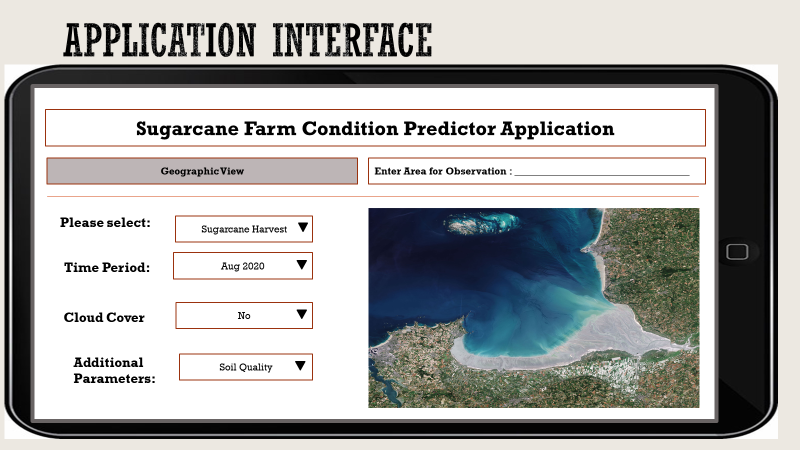


Figure 2 : Application User Interface with Geographical View

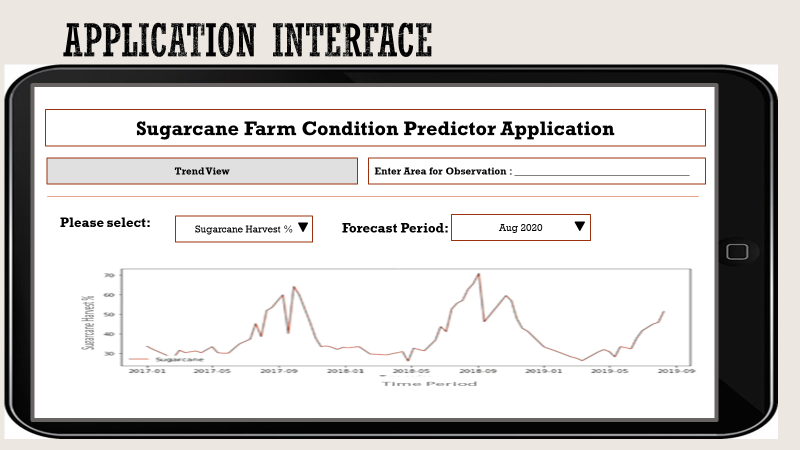


Figure 3: Application User Interface with Time Series Trend View

This will help farmers, government, banker and various business vendors take better informed decisions based on the predicted values.

# Methodology

Data exploration and Data processing

Image analysis

To understand image data, it is important to process it in machine understandable form. So, for the initial analysis, we converted image pixel data into .csv form for TCI, B02, B03, B04 band images. From this, it came to light that cloud and its shadow was posing serious impact on the pixel values and the values were anomalous when there was a cloud cover or cloud shadow.

From this analysis, we also understood that it would be extremely difficult to process entire data as there were many values involved. Therefore, first we aimed to mask the sugarcane field, remove cloud cover and shadow and color quantize the images. Then we decided to identify trend of vegetation for one sugarcane pixel.

Sugarcane Field Masking

As for this analysis, we are only interested in sugar cane farmland, we masked all the images using given masks for the respective area tiles.

In the following mask, black represents sugarcane area and white represents not sugarcane area. Thus, to apply this mask using bitwise and of OpenCV, we simply inverted all the masks and applied those masks to the sugarcane tile to identify sugar cane area. Example is presented in figure 5 which is a masked and after applying this mask on the respective image, we create tile presented in figure 6.

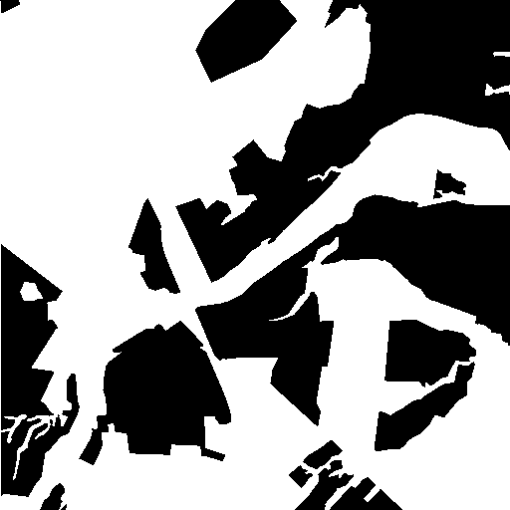
 

Figure 4: Sugarcane Mask Figure 5: Recover sugarcane area

Once we recovered the area of interest from all the given images, we turned to understand trend of sugarcane farm.

Time series of sugarcane farm data

In order to understand, whether there is anything meaningful in sugarcane farm pixel we wrote a python script just to gather data for one sugar cane field pixel. We run this analysis on band B02, B03, B04 and TCI image and got following observations.

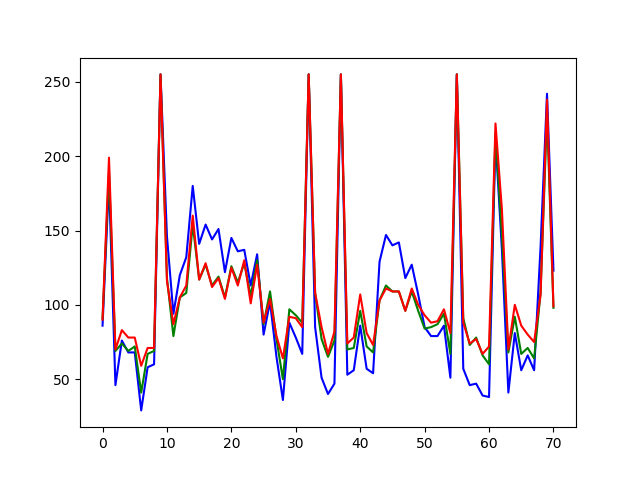


Figure 6: Time Series of a sugarcane land pixel in TCI images starting from Dec 16 to Aug 19

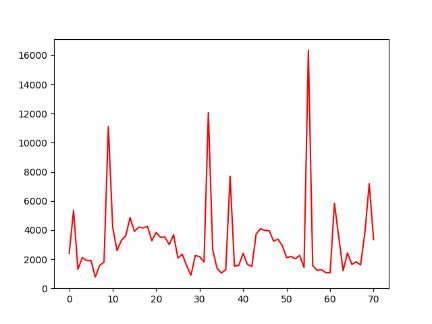
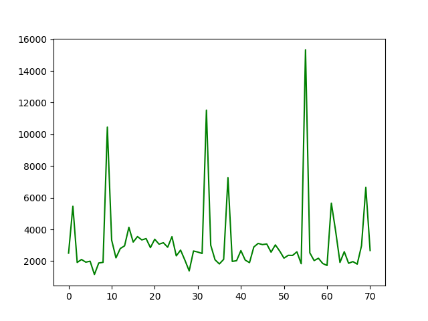
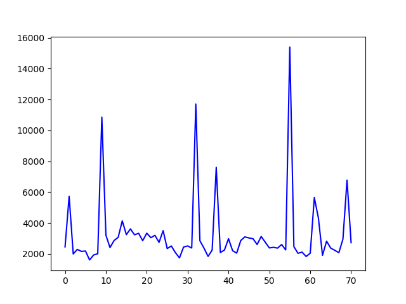


Figure 7: Time series of a sugarcane land pixel in B04, B03 and B02 images over three years

From this analysis it was very evident the present of cloud has affected the time series very heavily. Thus, the first step that we need to take is to remove cloud cover. For example, in figure 7, we notice some spikes in the data, representing that this value are not much in line with the expectation and the similar trend is also observed in the images provided in figure 8.

Secondly, seasonality was very clear, indicating that sugarcane has a distinct life cycle. From further analysis and with the use of metadata, it was clear that sugarcane planting takes place in March and April and harvesting occurs continuously for 5 months from June each year [1].

Cloud Cover Detection

In order to detect and mask cloud cover, we wrote a sophisticated python code which uses python libraries such as OpenCV, NumPy and pandas. We set a threshold to capture cloud pixels and masked them in the image so that impact on cloud cover could be minimized.

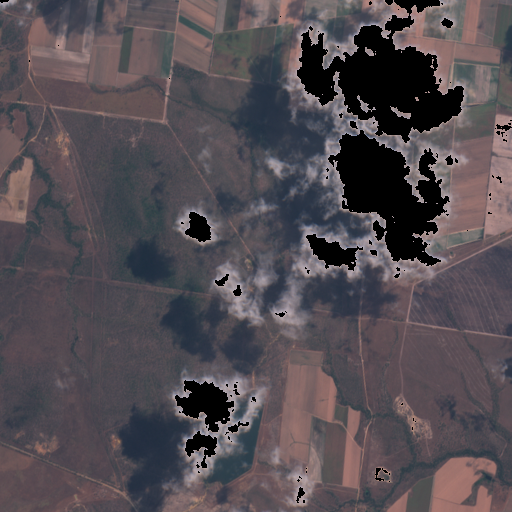
 

Figure 8: Original Image with cloud cover Figure 9: Cloud cover masked image

In the above images, we have the original image (figure 9) and the processed image (figure 10) where the cloud cover has been masked.

Cloud cover thresholds are upper bound = [180, 180, 180]; lower bound = [0, 0, 0]

In the same way, we also masked shadow pixels and processed all the image for the analysis.

Color Quantization

Color quantization is the process of reducing the number of distinct colors in an image. Normally, the intent is to preserve the color appearance of the image as much as possible, while reducing the number of colors, whether for memory limitations or compression. [7]

The idea behind this step was to use only one color for same age sugar cane crop. As a farmer plants crop at the same time for his land, there is no point of keeping track of all the pixel values. As our final aim is to provide information about various stages of sugarcane crop in the selected image using NDVI index, reducing number of colors made process more uniform providing improved accuracy over un-quantized image.

To implement color quantization, we used KNN algorithm to cluster the image data. Clustering is very useful to group similar values together and separate them from dissimilar values.

For implementing this step, identifying correct value of K was very difficult, thus decide value of K using a very simple script where we have considered the masked area to the entire area. If the masked area in relatively smaller, then we consider lower K value and if the masked area is larger, we consider bigger K value. The examples are presented with K = 12 in figure 11 and 12 and K=8 in figure K = 13 and 14.

After this step we again checked time series plot of each cluster and it shows very impressive trend to support the analysis that planting takes place in March and April and harvesting occurs continuously for 5 months from June each year.

Figure 10: Original Image Figure 11: Color quantized image at K = 12

Figure 12: Original Image Figure 13: Color quantized image at K = 8

NDVI

Normalized Difference Vegetation Index (NDVI) has gained popularity for its simplicity to predict crops health and it is highly used in precision agriculture. The Normalized Difference Vegetation Index (NDVI) is a standard band-ratio calculation frequently used to analyze ecological remote sensing data. NDVI indicates whether the remotely sensed target contains live green vegetation. When sunlight strikes objects, certain wavelengths of this spectrum are absorbed, and other wavelengths are reflected. The pigment chlorophyll in plant leaves strongly absorbs visible light (with wavelengths in the range of 400-700 nm) for use in photosynthesis. The cell structure of the leaves, however, strongly reflects near-infrared light (wavelengths ranging from 700 - 1100 nm). Plants reflect up to 60% more light in the near infrared portion of the spectrum than they do in the green portion of the spectrum. By comparing the ratio of Near Infrared (NIR) to Visible (VIS) bands in hyper spectral data, we can obtain a quick look at vegetation in the region of interest. NDVI is a normalized measure of the difference between reflectance at near infrared and visible bands of the electromagnetic spectrum. [2]

The formula for NDVI is:

**NDVI = (NIR - VIS)/(NIR+ VIS)**

1: NDVI calculation

To use this formula, for all the available image, we wrote a very sophisticated code which uses image of band B04 and B08 and calculates NDVI index for the image. These images provide detailed understanding of vegetation quality of that land.

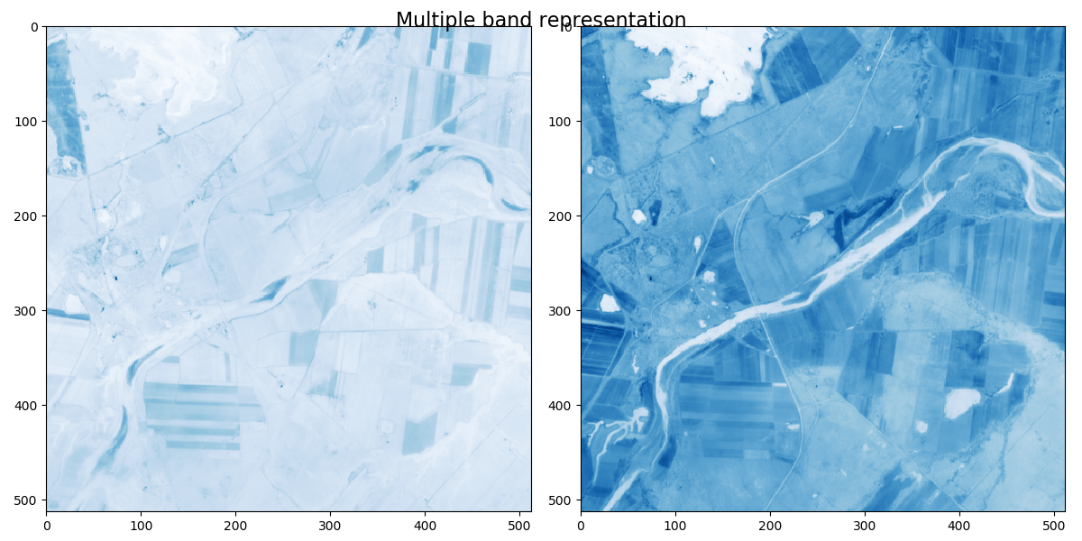


Figure 14: The band B04 and B08 images for NDVI

Different colors in the NDVI image represent different criterion about the status of the land.



Figure 15: NDVI band and value

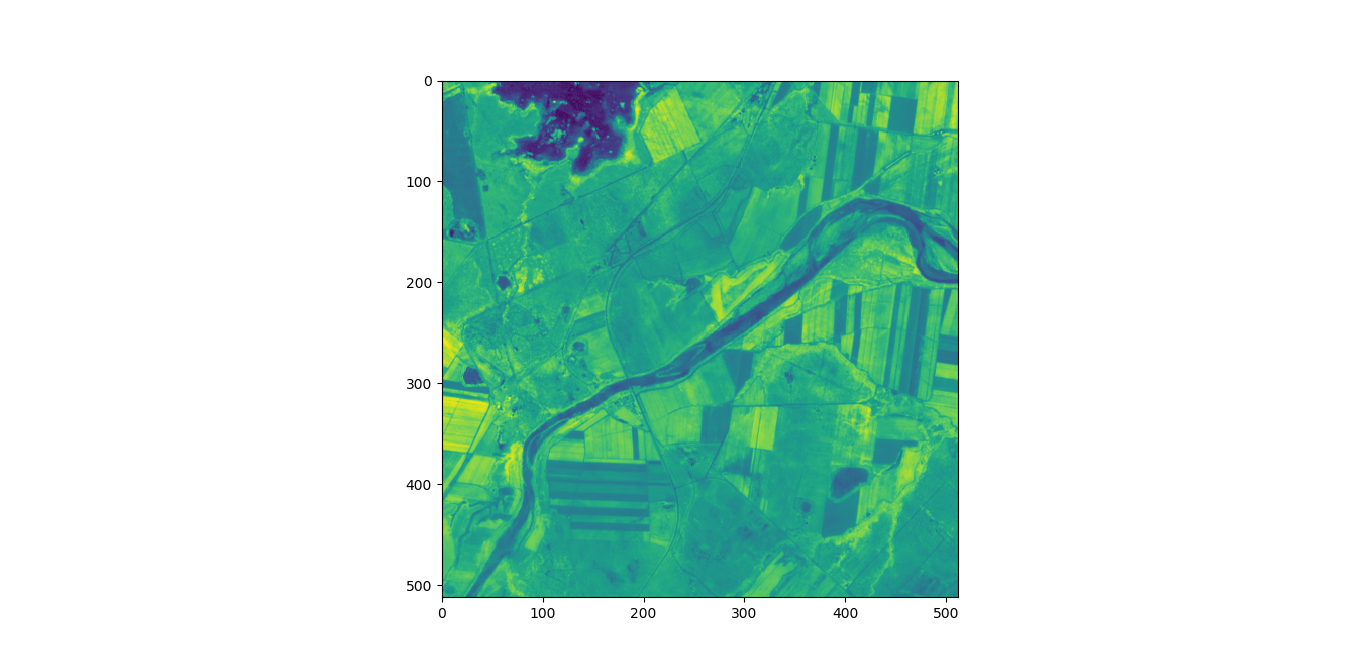
 

Figure 16: Original Image before NDVI mask Figure 17: Image with NDVI mask

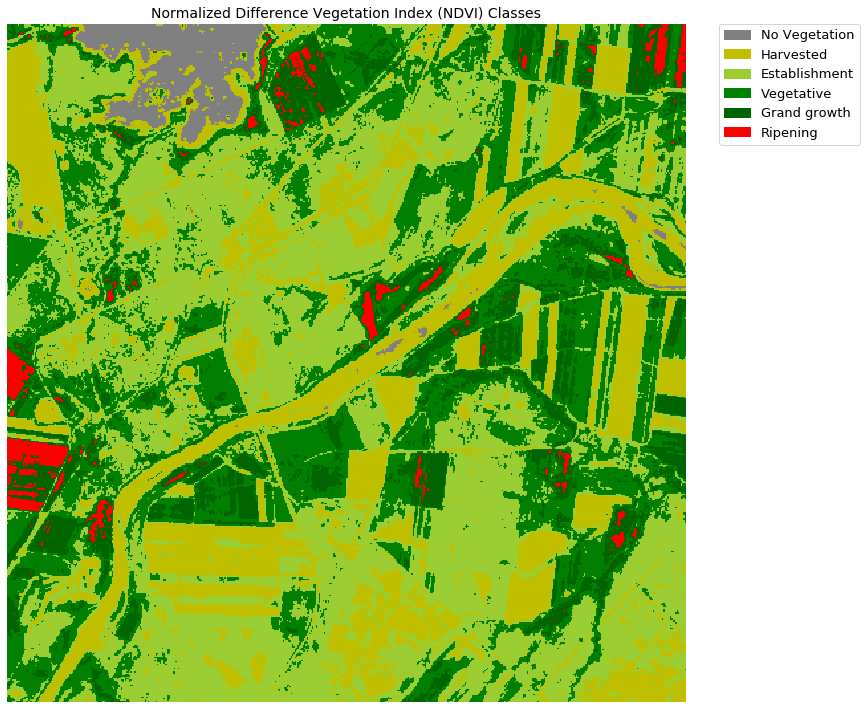


Figure 18: NDVI classified image

Ref figure 15, NDVI values between -1 and 0 correspond to non-plant surfaces that have a reflectance in the Red that is greater than the reflectance in the NIR. These could be surfaces such as equipment, water, or soil. Soil’s value is close to 0. Plant values range from 0.1 to nearly 1, and like stated earlier, the higher the NDVI value, the greater their density and health. [3]

From this analysis, we came to this conclusion that by considering NDVI, we divide life cycle of sugarcane into 4 stages:

* Harvested Phase: NDVI value between 0.0 to 0.25
* Establishment Phase: NDVI value between 0.25 to 0.35
* Vegetative Phase: NDVI value between 0.35 to 0.45
* Grand growth Phase: NDVI value between 0.45 to 0.55
* Ripening Phase: NDVI value above 0.55

After identifying thresholds for various stages, we plot figure 19 which successfully classifies various stages of vegetation in the given land. It successfully identified area without vegetation and identified recently harvested areas.

We use this information to understand various things about the sugarcane trend and make our algorithm learn from historic data.

Predictive model

To build a predictive model which can predict view of site for the given time frame, we are training our CNN model. We have attached all the 71 images available for all the areas and we are converting it into video data. As this video data is input to our CNN model. We hope that once everything is up and running, this model would be able to provide accuracy at least up to 70%. Now, the accuracy is not very high; therefore, we are working on optimizing its output.

To implement web functionalities, we are using Amazon Web Services and Google cloud services.

AOuth: We are using google cloud service AOuth to get user signed when they visit our website, so that we can track popularity of our website in various areas and track visit paid by users on our platform. We store all this data into DynamoDB so that it could be used for further data analysis. We also intent to offer data analytics in the scope of this site, DynamoDB works the best with Amazon Athena which was another added benefit.

S3 and DynamoDB: We are using S3 (Simple Storage Services) and DynamoDB as our primary storage. We have stored all the processed data into S3 buckets, and we have connected S3 to DynamoDB. We have stored information parsed from JSON files into DynamoDB. However, we haven’t used all the available feature from the metadata, we have only stored those features which are interesting for our analysis. We also store values of image pixels for some analysis.

Google Places API: We have used google places API to capture longitude and latitude of requested place so that we can quickly check if you have that area in our record using BigQuery on DynamoDB table.

Amazon Lambda: We use Amazon lambda to run our code virtually without any administration. This provides very high-level simplicity to make our webpage faster and provide output in no time. As lambda can activate thing on some event, to implement many functionalities we rely on Amazon lambda.

Amazon SageMaker: Amazon SageMaker is very sophisticated service that provide high opportunities for running algorithm of machine learning. It is also highly compatible with TensorFlow and Keras which will give us added benefit for using Deep learning and CNN to predict future outcomes.

Google App Engine: We use Google App Engine to host our website. As Google app engine is very simple and cheap, we used it.

At the moment, we are exploring more options and services from Amazon and Google cloud platform.

# Conclusion

Business Model

Once we train our CNN model, we will host the model on web platform to make it available for public use. Currently, we have already set up our storage for web plat form and we have tested few basic features, such as address input.

We provide our web visitor to enter location of the area he/ she is interested in getting more information about status of sugarcane land. This allow us to get location co-ordinates and check whether we have information about this land. If we have information about this land, then we directly update the dashboard on the website. From this dashboard user can select an image for desired date.

Once user selects image, we provide all the functionalities shown in figure 3 and 4. User can also ask us to provide information about the stages of crop on that land, which we will provide using NDVI index and color quantization.

User can also ask for time series model on particular sugarcane land and we will provide a sophisticated time series plot of various stages occurred on the land for the available time period.

User can also request future view of particular land and our model will be able to provide the image.

We also gather the harvested area data to inform biogas industry about the location and expected harvest mass as it will help them to contact the farmer for buying sugarcane waste product. To identify how much waste would be generated we intend to use external dataset that can provide information about in general expected cane waste in 1 ht area.

We can also identify crops in ripening stage and inform nearest sugar mills about expected produce in their work. To identify how much sugar cane would be generated we intend to use external dataset that can provide information about in general expected cane in 1 ht area.

# Appendix A – Project Management

Team Members

Our team comprises of four members who are data analysts and scientists and below is a brief introduction of their profiles and their individual contributions.

Kriti [M1] is a data analyst who is pursuing her masters in analytics at RMIT University. She has worked in various industries like pharmaceutical, workforce management, retail and property tech as a data analyst. In this project, she was mainly in charge of project management, technical research, business understanding, data understanding, preprocessing, manipulation and model building. She used Trello for Project Management and python to preprocess, analyze the data and build models.

Radhika [M2] is a data science enthusiast who is pursuing Master of Data Science at RMIT University. She has earlier worked as Software Analysts in education industry. In this project, she was mainly in charge of technical side of the project. She implemented proposed ideas to see whether ideas where serving expected results. She worked on data wrangling, modeling, evaluation and deployment of the project phases. She used Python, JavaScript, BigQuery, HTML and CSS to develop working protocol.

Both team members contributed equally from project kick off to data preprocessing, data exploration, brainstorming data modelling and analytics.

[**Kind Note**: We had started this project with 4 team members and two of them dropped the project in first couple weeks of run. Other 2 member are alumni of RMIT and due to their professional reasons, they were not able to keep up with this project]

Communication Tools

* Microsoft Teams (Official RMIT Account)
* Microsoft Outlook
* Trello for Project Management and deadlines reminder

Issues Encountered

* Clash in schedules
* Ongoing study and work which makes it difficult to devote more time to the project
* Version clashes for Python and trouble is loading libraries
* Data conversion related trouble while working with ArcGIS and QGIS software
* Resource allocation and task management

Agile Methodology and Lean Model

We had weekly catchups unlike the concept of daily standup in Agile we met every week and coordinated online over web meetings.

We covered following in every scheduled weekly team meeting

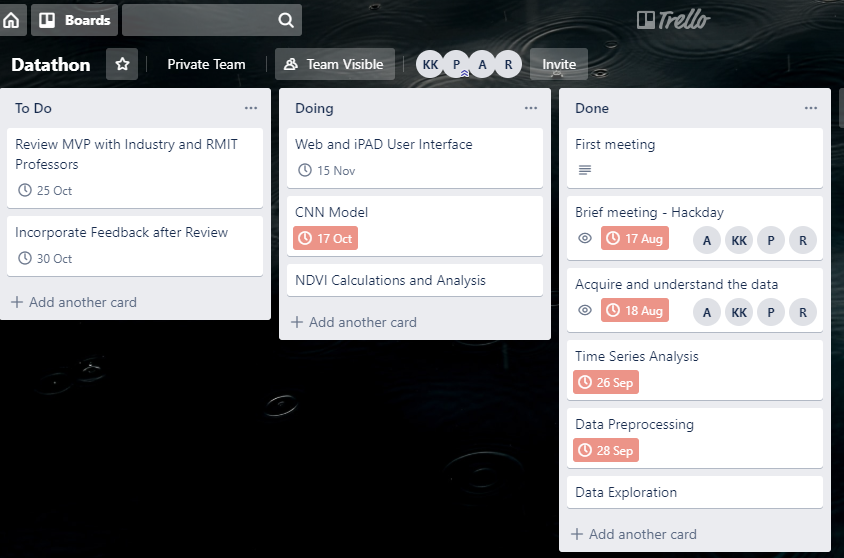
* Progress so far
* Plan for today’ s discussion
* Obstacles and challenges preventing progress
* New Ideas that can be implemented
* Technical glitches

Lean Model

We worked parallelly and tried to make a Minimal Viable Product (MVP) that we can test before we build the entire application. We started analysis using one image to see if the code is working and returning us correct insights. We started of writing small pieces of codes and then knitted them together for end to end solution. We made many quick attempts to building different algorithms and testing the output for better results. If we found out that a direction is not achievable, we evaluated and then tried different things.

Project Progress and Timelines

We used Trello for assigning tasks and gauging our progress and timelines.



Project Phases

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phase** | **Details** | **Who** | **When** | **Where** |
| Business Understanding | Understand Business requirements | Kriti Kumar (M1)/Radhika Zawar (M2) | 15 Aug 2019 | RMIT University |
| Define clear objective | M1 and M2 |
| Brainstorming on ideas | M1 and M2 |
| Data Understanding | Image Data Analysis | M2 | 20 Aug 2019 | RMIT University |
| JSON File Parse | M2 |
| Understanding and reviewing the data | M2 |
| Data Preparation | Creating Excel Data from JSON File | M2 | 10 October 2019 | RMIT University |
| NDVI calculation using image data | M2 |
| Data Manipulation | Time series analysis using data extracted from images | M2 | 17October 2019 | RMIT University |
| Predictive Modelling | M1 and M2 |
| Application buildup  Phase I | Web page with dashboard and location pick-up functionality | M1 and M2 | 19 October 2019 | RMIT University |
| Evaluation | Image Classification using CNN | M1 | 1 November 2019 | RMIT University |
| Application buildup | Web Application buildup | M1 and M2 | 3 November 2019 | RMIT University |
| Model Evaluation | Testing the model | M1 | 5 November 2019 | RMIT University |
| Presentation | Prepare a detailed presentation detailing the progress and findings | M1 and M2 | 7 November 2019 | RMIT University |
| Final Pitch in at Datathon Event | Final Presentation submissions at Datathon event | M1 and M2 | TBC | TBC |

Development Tools

We used different tools and programming languages for our analysis.

**Programming languages**

Python, BigQuery, JavaScript, HTML, CSS

**Tools**

Jupyter Notebook, Atom, Nodejs, Notepad++, Microsoft Excel, Tableau, Orange

**Software**

fmask, QGIS, ArcGIS Pro, Anaconda

**Third Party Services**

Python libraries, AWS, Google cloud services

**Project Management Tools**

Docker, AWS, GitHub

# References

[1]https://www.australiancanefarms.com.au/gallery, online platform, General information about sugarcane farming in Australia, last accessed 22-10-2019

[2]https://www.neonscience.org/calc-ndvi-py, NDVI related information, last accessed 21-10-2019

[3]<https://botlink.com/blog/discover-ndvi-and-its-valuable-uses-in-agriculture>, NDVI for sugarcane prediction, last accessed on 21-10-2019

[4]https://iopscience.iop.org/article/10.1088/1755-1315/149/1/012001/pdf, by Tri Muji Susantoro et al 2018 IOP Conf. Ser.: Earth Environ. Sci. 149 012001

[5]https://aws.amazon.com/products/?nc2=h\_ql\_prod\_fs\_f, cloud services, last accessed on 22-10-2019

[6]https://console.cloud.google.com/home/dashboard?project=decoded-petal-253712&\_ga=2.25029628.-778861555.1571734682&pli=1, google cloud services, last accessed on 22-10-2019

[7] https://www.pyimagesearch.com/2014/07/07/color-quantization-opencv-using-k-means-clustering/, color quantization, last accessed 15-10-2019