CASE STUDIES IN DATA SCIENCE

SATELLITE INTELLIGENCE

WIL PROJECT - GROUP 44

Yogesh Haresh Bojja (s3789918)

Nikhil Horakeri (s3828740)

Pankaj Rajendra Jagtap (s3789922)

Sagar Suresh Kulkarni (s3828982)

Contents

Introduction:

Sugarcane is a perennial grass which hence it contributes to the economy throughout the year. In terms of quantity, it is the largest grown crop. Australia contributes to around 33.5 million tonnes of sugarcane production [1]. It ranks second in the world in raw sugar supply. Sugarcane's contribution to the economy is around $2 billion per year. Each year there are around 378,000 hectares under sugarcane, out of which 95% is produced in Queensland [2].

There are many challenges faced in the Sugarcane sector, these challenges are faced by the farmers, mill industries as well as the bankers who provide loan to the farmers. There are many stakeholders like farmers, banks, sugar mill owners, etc. Being a farmer, he should be able to predict the health of his crop, so that he can get a better production, farmer should be able to gain insights into the comparison between his field when compared to other farmers' field. Banks should provide proper loan amounts to farmers by predicting their harvest production so that they can be able to return their loans.

Problem Definition:

Sugarcane is the source of Sugar and the current benchmark first-generation feedstock for efficient biofuel production. Sugarcane is one of the most exported products in Australia. More than 80% is exported as bulk raw sugar, which makes Australia the second largest Exporter in the world. For farmers who own a 2–5 acres field, the sugarcane continues to be the main crop. They are characterized by the continuous growth of sugarcane with and without rationing. For these farmers, the extension services provided by the sugarcane company are crucial to keeping sugarcane farming a viable activity.

Worldwide sugarcane is affected by a diverse number of viruses. Sugarcane mosaic disease caused by Sugarcane mosaic virus is economically the most important of all and is seen in almost all the major sugarcane-growing countries [4]. So, a farmer needs to understand the quality of sugarcane produced in his field. So, the farmer can compare his sugarcane produce quality with the other farms. The vegetation index and chlorophyll are some of the important features which define the quality of a crop.

From the perspective of a Banker, he should be able to decide whether he can provide a loan to the farmer based on land evaluation and the crop to be harvested. So in a precise way, the problem statement could be stated as follows: **To develop a web application that determines the yield and assess the sugar cane production of a specific agricultural land by using the satellite image and the land evaluation as a parameter**.

Proposed Solution:

The initial prototype was to use to first find the satellite image of the field on a particular date and find the field which is harvested using the edge detection and Kmeans clustering to cluster the pixels into a specific region.

For the next stage, the cloud detection concept was used to detect and remove the clouds and analyze the satellite image for further processing. The vegetation index and chlorophyll features were also included to define the quality or the health of the crop.

The Last stage consisted of getting the land evaluation prices of each of the plots from the Queensland website and using the shapefiles to get the plot shape and determine the yield and the price after harvesting.

Methodology:

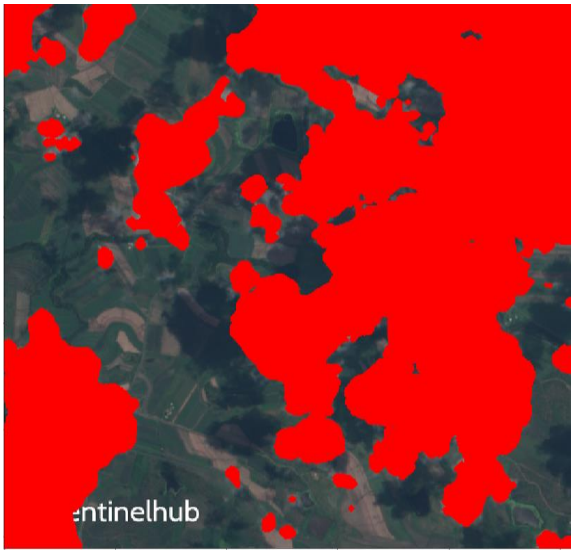
4a. Data Preparation:

Initially in the Datathon, we were provided with TCI images i.e. True color images and band images from B01 – B12 including B8A. These images were from date 2016-12-22 to 2019-08-09. There were in total of 74 TCI images for this period. There were in total 994 images including TCI and all the band images. We found out that originally there were 168 images present for the given period, but we were provided with only 74 images. Hence, there was more than 50% of the data loss. Taking this into account we decided to work with all the images. To download all these images, we decided to use Sentinelhub API. We had to provide exact tile coordinates which we want to download to Sentinelhub API in form of BBOX i.e. Bounding box. We were provided with coordinates of tile in the Datathon in the geojson file. With the help of BBoxFinder which is the online tool, we got BBOX of our desired tile for the given geojson file in wgs84/EPSG:4326 formats. By giving BOX to the Sentinelhub API we downloaded all the 168 images along with its 13 band images.

4b. Cloud Removal:

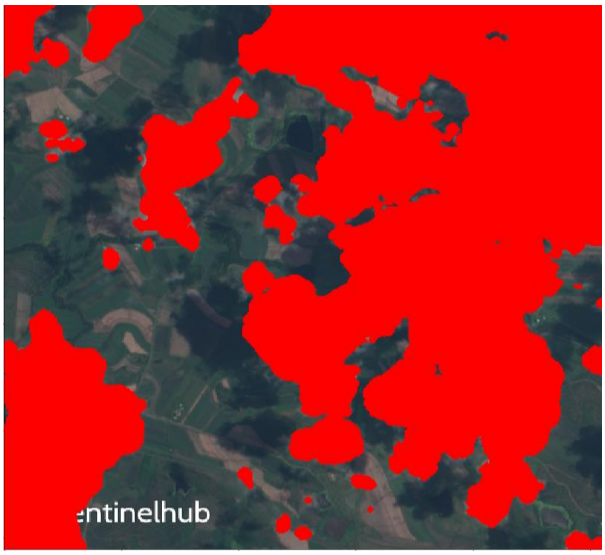
Several images in our data had clouds on it, which was a major hurdle for us to overcome. We used the S2Cloudless package of python for detecting the clouds from the image. Once we detect the cloud in images the part of the cloud was indicated in red color. The output of the cloud detection is shown below.

(TCI Image)



(Cloud Mask Image)

Once we detect the cloud in TCI images, we have converted the Mask image into Binary Mask i.e. white pixel depicts “Clouds” and black pixel depicts “No Clouds”. You can say that this Binary mask image is an image with mode “1”. Binary Mask created from Cloud mask is shown below.



(Cloud Mask Image)

(Binary Mask Image)

With the help of the binary mask, we can check for the given pixel if the cloud is present or not. If the cloud is present, then we have replaced part of the cloud from the image with the exact part of the image from the previous date. Dates have a difference of 10 days between them hence our assumption for this is for the given two dates fields are unchanged. The final output of the cloudless image is shown below.

(Binary Mask Image)

(Cloudless TCI Image)

4c. Shadow removal:

Shadow removal was another such hurdle for us. After comparing pixels of the image having clouds with the pixels of the image having no clouds, we found the value of the shadow affected pixels were dimmed averagely by 40%. Hence, we handled it by brightening the pixels affected due to shadow by 0.4 using the Python Pillow library.

4d. Creating Shapefiles:

After working on cloud images, we had to locate the fields present in the image. We found out the outline of the field with the plan number and lot number on the Queensland globe website. The plan number along with the lot number could identify all the fields uniquely. Referring to the outline of the field we created the shapefile for the tile in QGIS. The field outline along with the shapefile created is shown below.

(Field outlines)

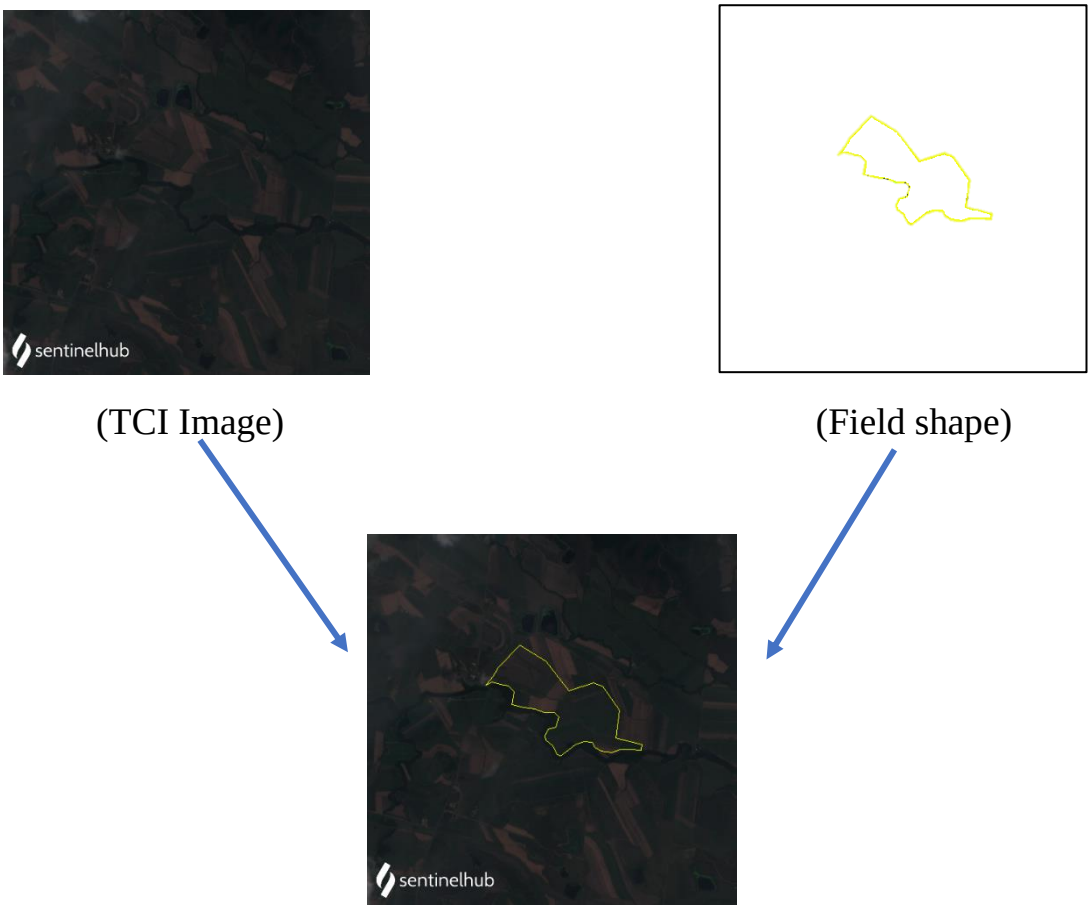
(Shapefile)

4e. Creating Geo Dataframe:

Queensland globe website also had the information on prices for each field shown above. So, for further analysis, we have included it in the shapefile. We have used the GeoPandas library to load the shapefile and work on it. We used this newly created column of field prices for further predictions and analysis.

4f. Superimposing plot shape on TCI image:

Whenever the user enters a date and plan number, we fetch the cloudless TCI image for the given date and field shape for the given plan number then we superimpose the field shape on the TCI image. The figure below can help us to understand the behind scenes.

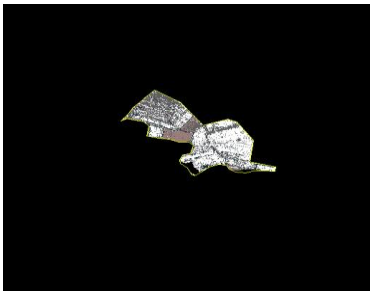


Hence, we get the superimposed images where the field is shown in the yellow outline. Therefore, we can do our analysis in this yellow outline region.

4g. Detecting Field for harvest:

The shapefile is superimposed on the satellite image to analyze the particular field. The Kmeans clustering and edge detection technique are used to detect the fields which are ready to be harvested. We have used the hue value of the green pixel to filter the pixels of the corresponding crops to be harvested as the hue considers the intensity value of the green pixel.

(Superimposed shapefile)



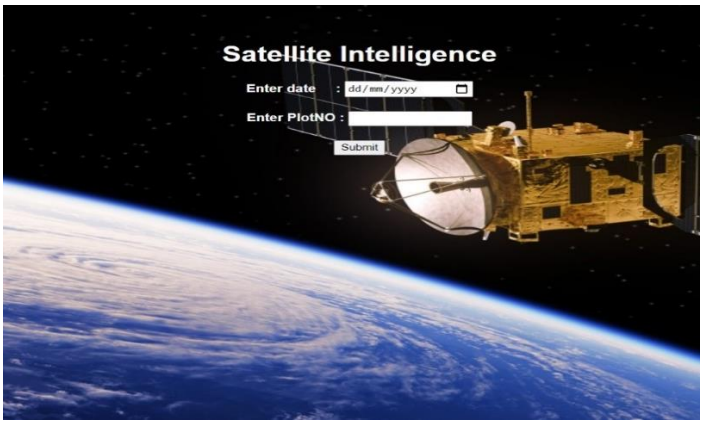
(Crops to be harvested)

We have converted the hue green pixel values which fall in the range (100 to 140) to white pixel, to differentiate the crops ready to be harvested.

4h. Methodology for use case1 (Banker):

As a banker, I want to be able to offer loans to farmers based on the value of their crops when harvested.

For this use case the web app accepts the date and plot_PlanNo of the specific sugarcane field as the input as shown in the below figure:

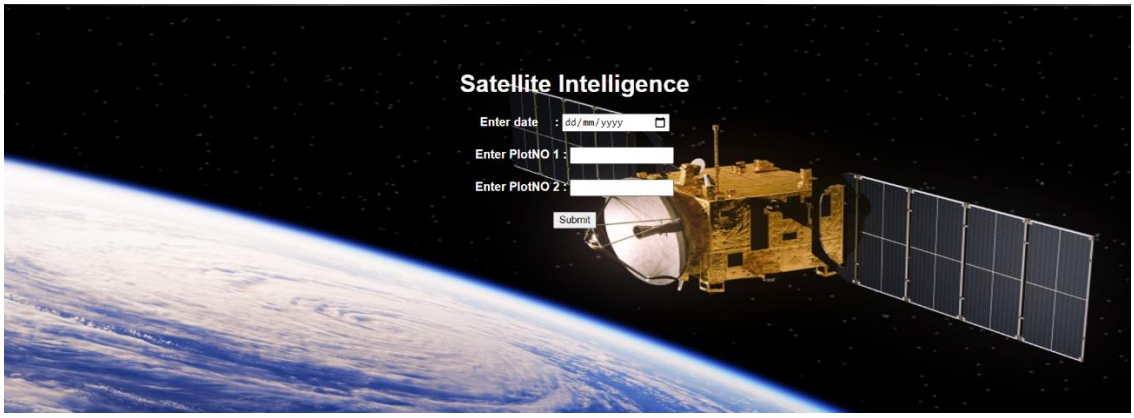


- We retrieve the specific satellite image based on the date mentioned by the user.
- We use the cloud removal technique which is mentioned in 4b.
- Once we obtain the cloudless image, we detect the field to be harvested using the technique mentioned in 4g.
- Using the plot shape, we get the area of the field in hectares and we use the details to calculate the yield of the land.

4i. Methodology for use case2 (Farmer):

As a Farmer, I want better visibility into how my crops compare to the other farmers in the area.

For this use case, the web app accepts the date and two plot_PlanNo of the specific sugarcane fields to be compared as the input as shown in the below figure:



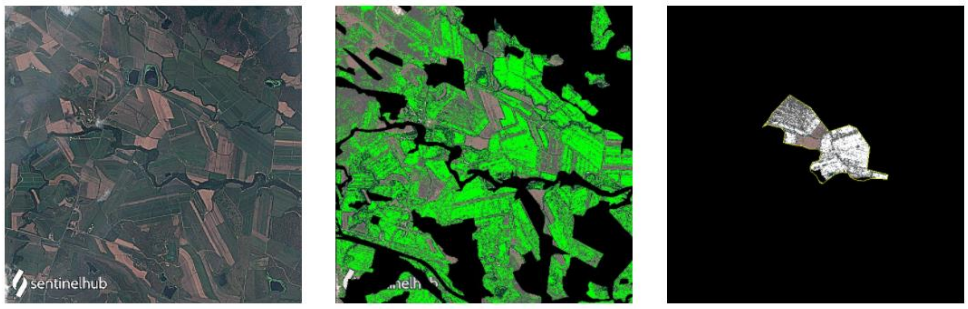
- We retrieve the specific satellite image based on the date mentioned by the user.
- For the farmer to compare the crops we choose to use vegetation index as the feature. The vegetation index defines the quality or the health of the crop.
- Using the plot shapes, we get the area of the field in hectares and we use the details to calculate the % of crops in the field which is in a good condition and can be used for harvesting.

Significance of result:

5a. Output for use case1 (Banker):

Considering the user selected a specific date and Plot_PlanNO(20-RP736247), below is the outcome of the analysis:

Images



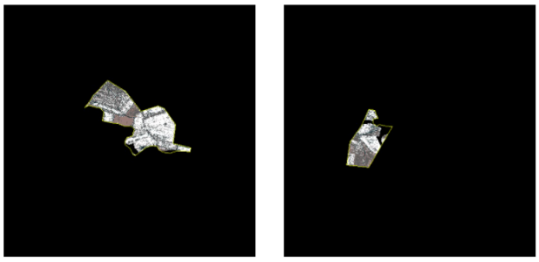
The land evaluation for plot 20-RP736247 is 7,40,000$
The area of the field is 1.2764 HA
The price per HA is 5,79,755$ per HA
The area of land in HA which is ready to be harvested is 0.668 HA

Based on this evaluation a Banker can decide whether he can offer loan to the farmer.

5b. Output for use case2 (Farmer):

When the user enters two specific plot_planNO, below is the outcome of use case 2:

Images



Plan No	Total Area in HA	Pixel Matched	percentage of Land ready for harvesting
20RP736247	1.2764	7765	60
2-SP107115	0.5507	4402	79

From the above comparison based on the vegetation index and the percentage of land ready for harvesting we can conclude that Plan no 2-SP107115 has a better yield compared to the field with Plan no 20RP736247.

Validation of data:

Even though the plan numbers and field shape we have got are from Queensland globe we tried validating it by google maps. The approximate estimates we calculated are based on the land evaluation prices which were retrieved from the dataset downloaded from the Queensland Website. The dataset had the unique plot and plan no to identify each filed and also had the corresponding land evaluation for that specific field.

Future improvement:

For future scope, we are thinking of extending our analysis from the Proserpine region to the whole Queensland region. At present, our application can allow farmers to only compare the harvest outputs with other fields but in the future, we are thinking of embedding other useful values extracted from band images like chlorophyll level present in the crops, Nitrogen and moisture level in the soil for the farmers' interface to compare it with the other fields.

Ethical, social consideration and fairness:

Ethical issues should be carefully considered in data science Project. The tool depicts or determines the approximate estimates of the farm yield which are largely based on the analysis of the satellite images. There can be risk of misuse that can use financial

harm to the farmer or the banker. Privacy is always important when it comes to any data science project. In this case the farmer may disagree for the data to be collected about the harvest details of their farm. So careful consider the needs of farmer's privacy.

We can measure fairness in the context of our WIL project as it involves satellite image processing of sugarcane fields of specific areas as we are dealing with land evaluation. But most of the images need cloud detection as these clouds overlay shadow on the fields. If the clouds are not detected accurately this might affect the accuracy as the image processing model will not be able to get the green pixels correctly which might lead to incorrect land evaluation. We can measure fairness by using a metric to measure how much the clouds affect our prediction. So we could have a confidence factor to display by which we can say we are 68% sure that land evaluation will be specific value(XYZ) because there are clouds in the image. In this way there will be no unfair land evaluation for the farmer.

Project Management:

We all have contributed equally sticking to respective tasks given and completing them. We planned in the excel sheet of all the tasks that are to be done along with the deadline and as the tasks got completed, we checked them as "done". In the initial phase of our project development, we took quite a lot of time as we were new to each other and due to the pandemic we couldn't meet, Therefore, we used Video conferencing as our major tool to connect in person and to know each other as well as discuss the tasks. Video conferencing was done at least once a week and, in some cases, when the tasks were a bit confusing and difficult, we had to do multiple of them. Each member had a virtual "stand-up" to discuss whether the allocated task was completed or not, if completed he would share more ideas of improvising it and if not completed then he shared the obstacles experienced. During these hard times in pandemic whenever restrictions permitted, we met in a group for discussion.

Communication tools:

- Microsoft teams
- Outlook
- Trello for deadline reminders

Programming languages:

Python,HTML,CSS

Tools:

Jupyter Notebook, Notepad ++, Qgis

Python Framework:

Flask

Python Libraries:

Pillow, OpenCV, geopandas, numpy, shapely

Members	Contribution
Yogesh Haresh Bojja (s3789918)	25%
Nikhil Horakeri (s3828740)	25%
Pankaj Rajendra Jagtap (s3789922)	25%
Sagar Suresh Kulkarni (s3828982)	25%

Phase	Details	Who	When
Business Understanding	Understand Business requirements	Nikhil,Yogesh,Pankaj,Sagar	05 - Aug to 17- Aug
	Define clear objective	Nikhil,Yogesh,Pankaj,Sagar	
	Brainstorming on ideas	Nikhil,Yogesh,Pankaj,Sagar	
Data Understanding	Image Data Analysis	Nikhil, Yogesh	Aug 18 – 31 Aug
	Understanding and reviewing the data	Sagar, Pankaj	
Data Preparation	Extracting the Plot shapes	Yogesh	Sept 1 – Sept 7
	NDVI calculation using image data	Pankaj	
Data Manipulation	Detecting and removing the clouds	Yogesh, Pankaj	Sept 7– Sept 14
	Edge detection and field comparison using vegetation Index	Nikhil, Sagar	
Application buildup Phase I	Web page creation	Nikhil, Sagar	Sept 8 - Sept 12
Evaluation	Determining the harvested field using Kmeans	Nikhil	Sept 13 - Sept 18
Application buildup	Web Application buildup	Sagar, Nikhil	Sept 19 – Sept 23
Model Evaluation	Testing the model	Pankaj, Yogesh,Sagar,Nikhil	Sept 23 – Sept 28
Presentation	Prepare a detailed presentation detailing the progress and findings	Nikhil,Yogesh,Pankaj,Sagar	Sept 28 – Sept 29

References:

[1]"Sugarcane", *En.wikipedia.org*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Sugarcane. [Accessed: 16- Oct- 2020].

[2]"Cane", *Queensland Farmers' Federation*, 2020. [Online]. Available: https://www.qff.org.au/farming-in-qld/cane/. [Accessed: 16- Oct- 2020].

[3]"Sugarcane | Plant Health Australia", *Planthealthaustralia.com.au*, 2020. [Online]. Available: https://www.planthealthaustralia.com.au/industries/sugarcane/. [Accessed: 16- Oct- 2020].

[4] *Sugarcane - an overview | ScienceDirect Topics* (2020). Available at: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/sugarcane (Accessed: 16 October 2020).