GeoRoots

WIL Project Report

A WEB APP MODEL TO PREDICT SUGARCANE YIELD
BASED ON DATA PROCURED BY SATELLITE IMAGES

CASE STUDIES IN DATA SCIENCE

GEO ROOTS – LOAN ESTIMATION
USING SATELLITE DATA ANALYSIS

Group 57: RMIT UNIVERSITY

Ayesha Hojage (s3802865@student.rmit.edu.au)

Chandrakant Prajapati (s3797785@student.rmit.edu.au)

Jordan Mathew (s3778812@student.rmit.edu.au)

Komal Mehta (s3795392@student.rmit.edu.au)

Samnit Singh (s3804007@student.rmit.edu.au)

Contents

1) Introduction

Australia is the 2nd largest exporter of raw sugar producing 4 million tonnes of sugar each year. Sugar Industry in Queensland has high profitability, and it has a great influence through the government agencies and their policies which affects the global market price and weather conditions. The total revenue generated by this sector is 2.5 Billion dollars every year and this industry employees about 48000 people. Apart from being used in sugar industry, sugarcane finds its application in medicinal applications and ethanol production. The increase in demand of sugarcane is expected to reach 200 million tonnes by early 2024. This makes it critical to provide proper finance cover to the sugarcane community to meet future demands.

The major concerns for this industry are faced by farmers to get their loan approved for the land they own, but bankers which provide them loan need some statistics regarding crop yield and its harvest. For this, Data Scientist plays an important role to propose an intuitive solution for serving an intermediate layer between bankers and farmers.

Here, we are going to present a business solution for bankers which could help them provide loan to farmers based on crop yield estimates. For this , we will be designing a web application, in which we would be providing some functionalities like calculating NDVI (Normalised Difference Vegetation Index) based on the polygon selection of certain area in the map depicting harvest per yield along with the health of crop which will determine yield in future based on which an estimated loan amount will be given to banker which will help them determining the amount of loan to be sanctioned..

Now, let us explore through this problem scenario and its business solution in the next section.

2) Problem Scenario Description

Few decades back, there was a deregulation in banks in 1980's regarding the crop harvest and its maturity rate which poses a high risk for Banks to provide loan to farmers. Crop health was not taken into consideration for loan estimation which discourages banks from sanctioning their farm loans. Also, they lack the necessary information regarding crop harvest.

Also, crop like sugarcane takes a huge amount of time for harvesting and cultivation of 9 to 16 months affected by various weather conditions and many other crucial factors. The Sugarcane farming industry faces financial debt, and they have least accessibility to funds that can be utilised for their living. Because of these reasons, financial dislocation began to spread beyond the 'farm gate' as some producers were forced to seek credit from businesses as their accounts were frozen by the lender.

Moreover, Government plays an important role in generating policies and regulations for both banking and farming industries. Considering these policies government in Queensland come up with a community group known as "Cane - Growers" which helps Sugarcane community a lot. Also, it is seen sugarcane growth makes a huge contribution to the Queensland economy, which includes upstream industries supplying goods and services to cane growers, on over 4,300 cane farms across the state, and also downstream industries that transport, process and market the sugar that is produced.

The problem can be clearly defined through a given scenario which depicts that statistics of the crop harvest would help the bankers to estimate the loan amount. Every project manager of the industry come up with one or the other solution proposed to enhance its production lifecycle. Similarly, when loan providing team is focusing on providing a certain amount to farmers they will look on to some factors like good harvest and yield which also

determines the good health of the crop. As we know, crops like sugarcane requires intensive labour and skill along with raw material banks should be considerate of the collateral factors that a farmer can repay through. To pitch in and solve the problem of the bank reluctancy and farmers debt issues, we as data Scientist took the help of Satellite images through Landsat 8 of the farmland and utilise their techniques to resolve this problem.

According to the data retrieved from LandSat-8 Satellite images through Image Processing, a predicting model has been made to estimate the crop yield based on which loan estimation and loan approval process is easily accessible to banks.

Figure 2.1 (Contribution of sugarcane on economic growth of Queensland)

The above statistics depicts the contribution of sugarcane to the economic growth of Queensland. The growth is having its backward linkages to its own suppliers (e.g. of fertiliser, herbicide, etc.) and its forward linkages to mills, refineries and the sugar terminals which are exporting the sugar. To boost the performance of this economic upliftment, banks are contributing to it by providing the support to the entire economy. For this the fluctuating yield estimation should be done which is the major challenge for the sugar mill industry. One of the major hurdles we got in estimating the yield is cloud cover which is not penetrated by Satellite when the weather is rainy. Also, the various climatic factors inculcate gaps in the prediction and estimation of yield of sugarcane. Factors like variation in temperature, changes in soil moisture and excessive rainfall or less vastly effect the estimation of harvest.

3) Proposed Solution

"Georoots" is a prototype designed to demonstrate the loan estimation based on the yield of the harvested crop also monitoring of crop yield during the crushing season. The dashboard application mentions these major areas:

- Vegetation of the crop area of Queensland on map
- Predicted crop yield and its estimated revenue over the span of 10 years
- Loan Calculator Based on NDVI Index
- Demographic information of the harvested crop

This tool can be scaled further to include features like land risk modelling by including information about the land area and condition. Currently the loan approval or the amount estimation is limited to a particular case which just adds up to the operational cost of the bank and lengthen the loan approval period causing delay in providing funds to the farmers. Our app will aim at cutting the delays and easing the process both for farmers and the financing institute.

Our tool helps in speeding up the process of loan approval with the help of estimating the future yield of the farmer based on the farm's location. Also, our tool will provide user friendly interface for loan processing for the banks and lays a hazel free path for farmers to validate their yield production.

The yield estimation of sugarcane is done by satellite image processing of the fields whose data is available as a part of **Copernicus programme** from which is the European Union's Earth observation programme coordinated and managed by the European Commission. The tool will estimate the future yield of the farmer and will help the banks to provide the appropriate amount that can be approved based on the current market value. This will be done using machine learning techniques with the help of NDVI (normalized difference vegetation index). NDVI is a simple and impactful index which is used to quantify green vegetation. This index helps in identifying whether an area contains live green vegetation.

Let us see how the estimation is done through a certain methodology.

4) Methodology: -

4.1) Dataset Description

The data provided was generated by the Sentinel 2A satellite which is a remote sensing platform and is capable of capturing images of different wavelengths. The datathon dataset contained satellite images of Queensland's sugarcane growing area just outside of Proserpine. The dataset from the first phase contained images of a small region of the entire area. Along with this, a timeseries folder contained 994 image files of that particular area. The images were taken approximately 10 days apart by the Sentinel 2A satellite and provide imagery with a 10m per pixel resolution. The images provided were given in 12 bands and, the timeline of these images range from December 2016 to August 2019. For the analysis, the TCI images included all the different colour bands (blue, green, red) together and were used as the proportion of different colour bands that can be compared with each other. The images were provided in timeseries as each tile is time stamped by the date of capture. Along with images there was a json file providing meta data about the conditions of the capture, and its location in lat/long.

Example of Sentinel-2 Images:

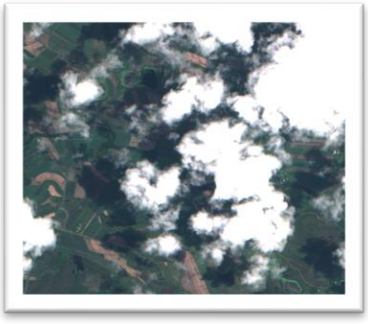

Figure 4.1: Land with Cloud cover

Figure 4.2 Land that is harvested

The clouds in images have been masked using s2cloudless, which is a machine learning for calculating cloud masks on Sentinel-2 satellite images. This package uses a Light Gradient Boosted Machine (Light GBM) machine learning model to quickly identify clouds in an image, based on spectral bands of Sentinel-2 satellite imagery. Some reverse-engineering was a requirement to modify satellite imagery into the "Sentinel Hub" format expected by s2cloudless and cloud masks were over-lapped onto different images to determine which pixels were to be hidden by clouds.

Figure 4.3: Corresponding “cloud mask” from s2cloudless

NDVI Index (Normalized Difference Vegetation Index)

NDVI index helps in inspecting fields and crops using satellite imagery. This index acts as an indicator of a plants’ health based on how a plant reflects different light waves. For understanding the health of a plant, comparison of absorption and reflection of red and infrared light waves is a must. And, the NDVI index does this job for us. The formula of the NDVI is given below:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

where, Red and NIR stand for the spectral reflectance measurements acquired in the red and near-infrared regions, respectively. By design, the NDVI varies between -1.0 and +1.0. NDVI values -1 and 0 refer to areas that contain snow, water, sand, and stones. NDVI values for plants range from 0 to 1. The NDVI index is one of the most popular indexes, and it also has a main advantage, that is, the high resolution of images with data from satellite imagery.

4.2) Tools and resources: -

The data used for this solution is taken using Landsat8 from Google Earth Engine and Sentinal2 comes under **Copernicus** program. Additional data for agricultural data is taken from Australian Department of Agriculture, Water and Environment. Whole software development and demo was done in Python 3.7 and R studios. Code for creating cloud masking was provided in datathon phase -1 and phase 2 for heads up.

4.3) Working: -

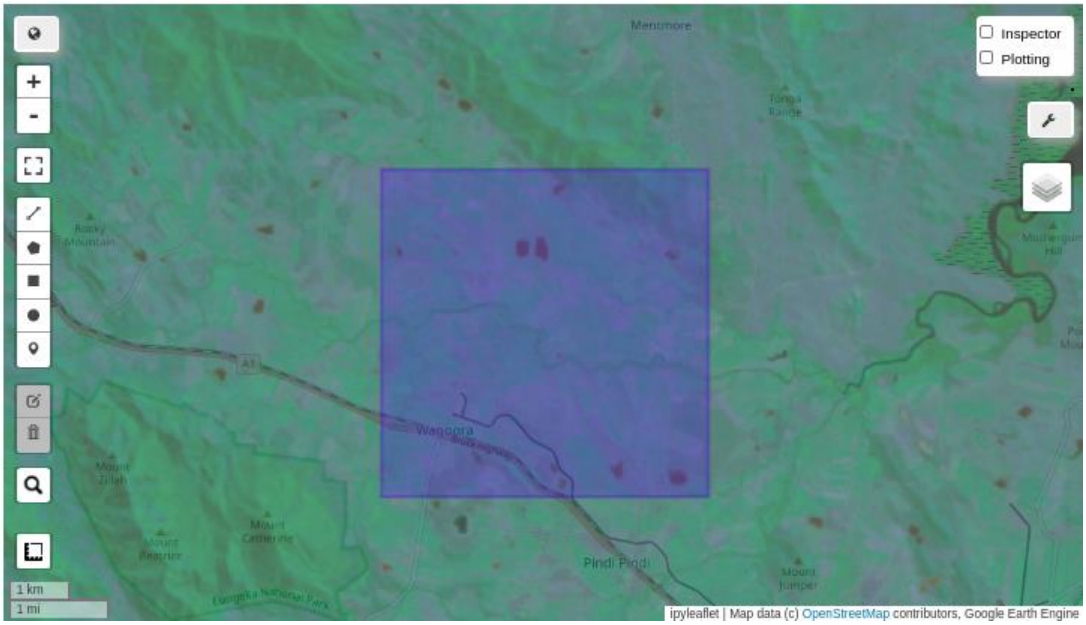

Figure 1.2 (Polygon selection on area of interest)

The tool can be used to select the area after searching the address one can manually outline the area of land in hectors for which the banker want to estimate the loan Banker can check history by typing the date rang then using the Landsat8 sentinel 2 data. NDVI, area of land and their respective yield for past 10 years will be shown.

To estimate the NDVI index Landsat8 Band4 and Band 5 is being used as the formula for calculating NDVI index is given as follows.

NDVI = (NIR- RED) / (NIR+RED)

Here the *RED* represents the red spectrum & NIR represents near infrared spectrum. The images of Band 4 and Band 5 is given as follows.

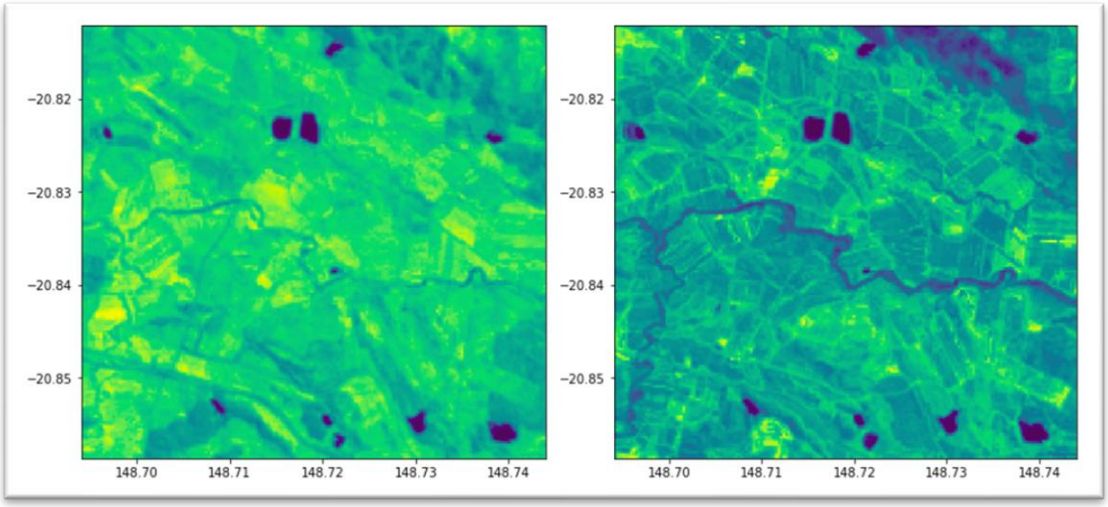

Figure 1.3 (Landsat8 data for Band4)

Figure 1.4 (Landsat8 data for Band5)

From the above plot, Band 5 is informative as the boundaries of the farmland is more clearly visible and descriptable.

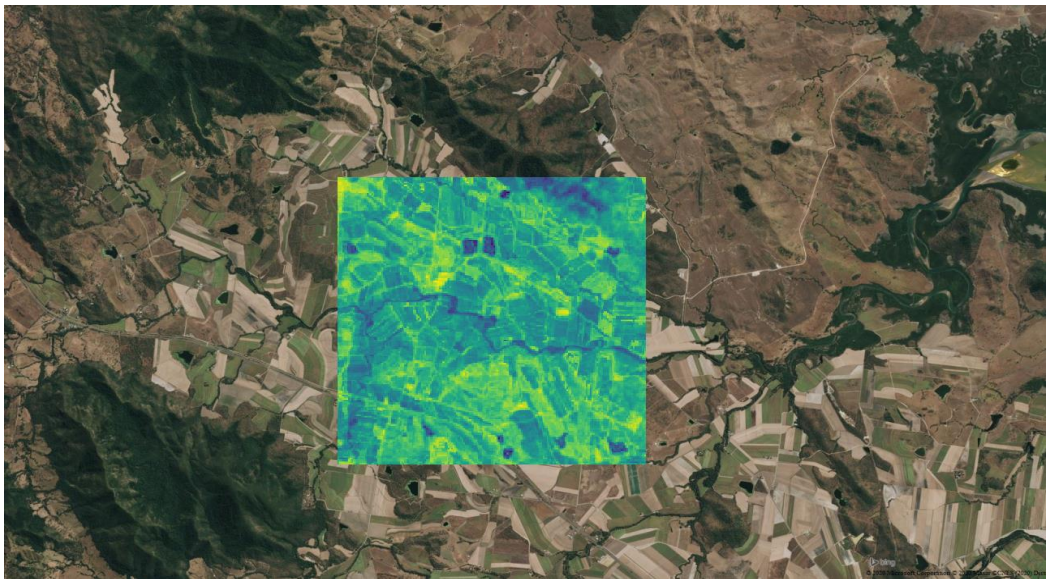

Figure 1.5 (Visual representation of NDVI index of selected land area)

The final NDVI data is overlapped on the real map to give the overall idea about the NDVI visualisation. Here the blue is representing water and yellowish green region represents land with high vegetation index.

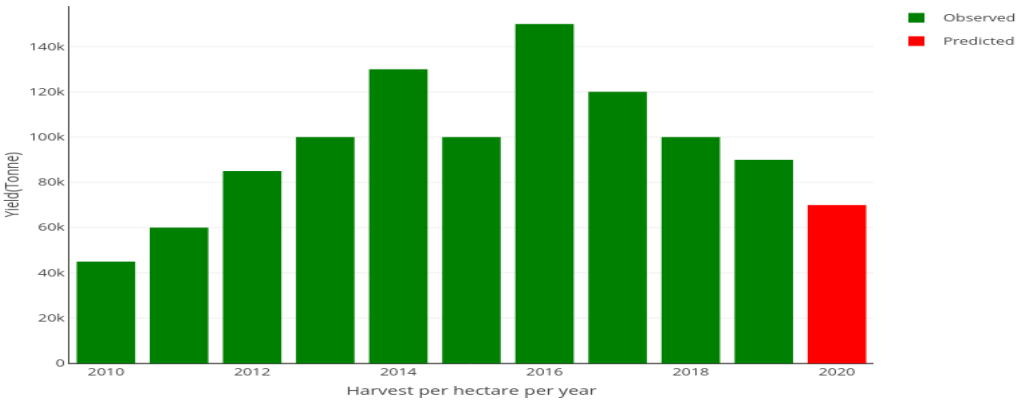

Figure 1.6 (Yield per year)

Using the Agriculture dataset from the Department of Agriculture Water and Environment data and from past year gives us the rough estimate of the past yield as well the current predicted yield which help banker determine the loan easily.

Finally, the dataset which contains NDVI index and demographic information of the land such as country, state, region and season. Area of the land in hectares and its corresponding Yield is also noted and all together these values are passed to the Linear regression model to estimate the loan amount which can be sanctioned to the farmer with low risk.

5) RESULTS AND CONCLUSION

Our Business Problem was mainly focused on providing the loan estimation for the agricultural land. The tool helps in estimating the loan by the through assessment of land using Satellite Imagery which help the bankers to understand the condition of land based on past Satellite data and its corresponding demographic provided by Landsat- 8 Satellite.

This also helps to eliminate the field trip to land resulting in fast loan approval. By considering these factors, the collateral by farmers is also clearly describe to Bankers. The prototype also provides the prediction of yield and NDVI Index which determines the estimated low risk loan amount which can be sanctioned to the farmer.

As we can see from below graph, the NDVI index of selected agricultural land based on which we can say that the land is fertile and estimated yield will be good. This good estimation is used by bankers to analyse how much they should grant loan to the farmers.

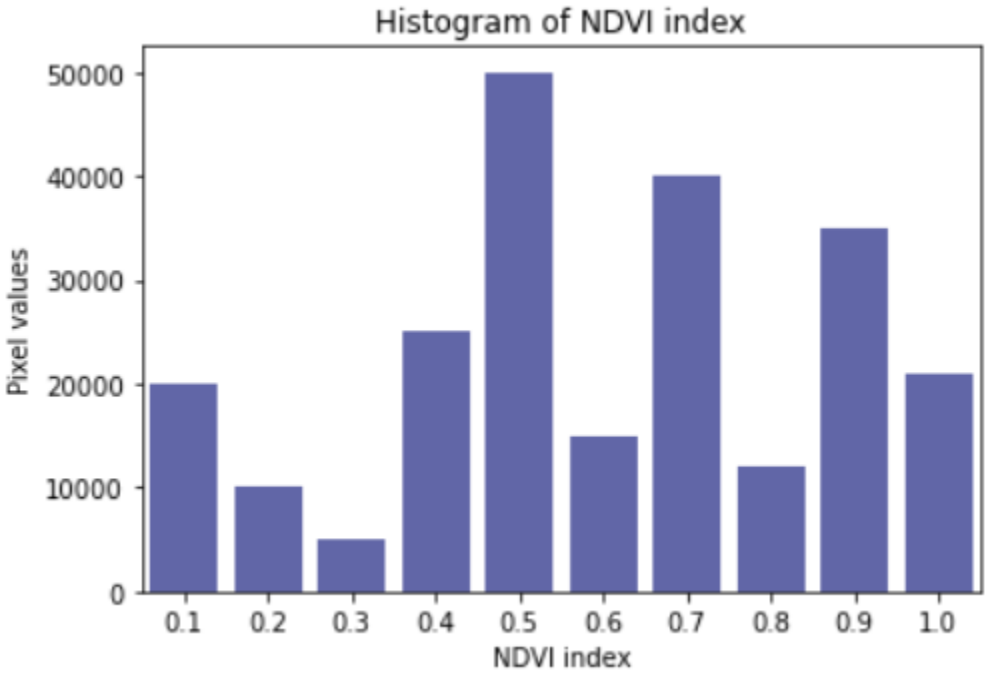

Fig 5.1 NDVI Index of selected polygon

6) Project Management Plan

To track the progress of the team, Microsoft Excel Spreadsheet was used. The sheet comprised of different phases of the project. The phases were included setting up the tools necessary for an effective communication between the team, the initial research and hypothesis analysis, understanding the data, designing, modelling, presentation, front-end development, and Video creation for presentation. Team has used the Agile methodology to cover these tasks, including milestones.

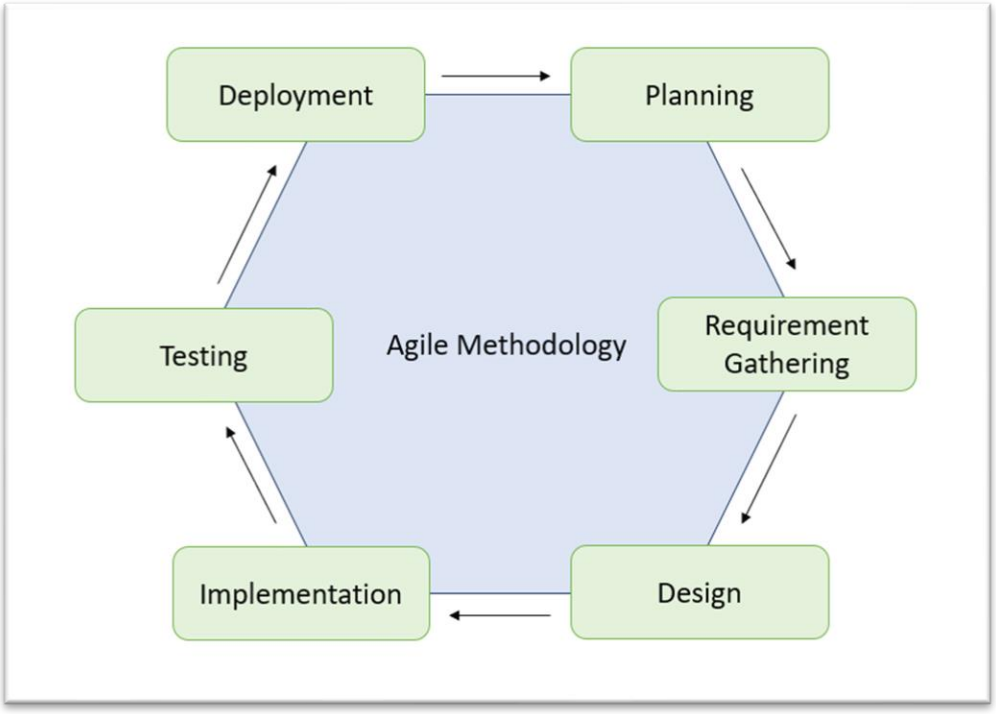

Figure 6.1: SDLC Agile model which was followed for this project.

As the team comprised of students with diverse backgrounds and skills, we were able to prioritize the tasks according to each team members strength and prior experience. While constructing a project plan, it is necessary to understand the time taken by every task, along with its dependency with any other tasks. Considering these factors team finalized a high-level plan. This plan helped team to work on their tasks, along with an equal share in the overall work.

Apart from using Microsoft Spreadsheet, the team used different communication channels to track updates of the project. This made it easy for team to communicate with each other whenever needed. In the initial period of the project, team decided to create a WhatsApp group to have short conversations and daily updates. Within 2 weeks the team set up a Slack Channel as a communication medium to share all the links, discussions related to tasks, any bottlenecks, helping other team members, discussions for improvement of any tasks.

Team used a combination of Eclipse to Set up Eclipse IDE for developing Python batch processes. Team also decided to setup a meeting on Microsoft Teams every Monday and Thursday Morning to understand the working of the project. Each team member would discuss how did they work on a task and how did they deal with the problems they faced while working on the tasks. These meetings gave each team member an opportunity to share the progress they were contributing to the team.

After every meeting a Trello board was maintained to understand to what's being worked on, who's working on what and to-do tasks. Trello board helped us to maintain keep points of the tasks needed to be worked on. Overall, each team member contributed equally for this Project.

Student ID	Student Name	Contribution (%)
s3797785	Chandrakant Prajapati	20
s3804007	Samnit Singh	20
s3778812	Jordan Mathew	20
s3795392	Komal Mehta	20
s3802861	Ayesha Hojage	20

Figure 6.2: Team Contribution

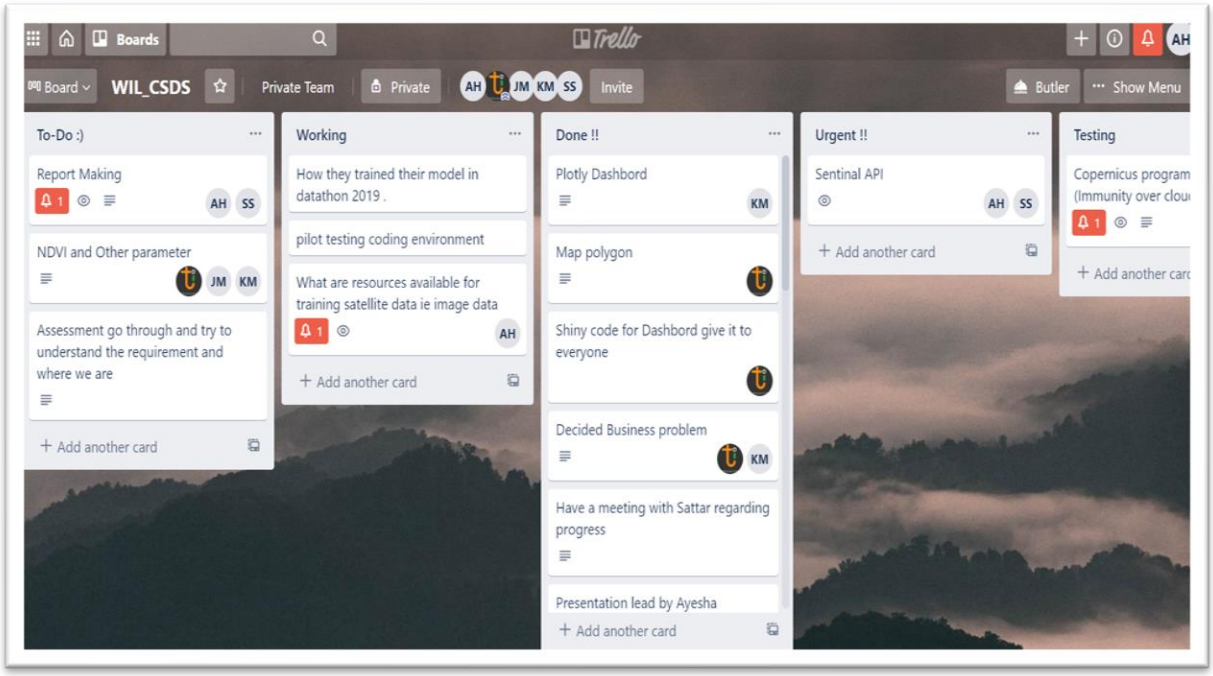

Figure 6.3: Trello Board

The final share of team member contributions can be seen in below:

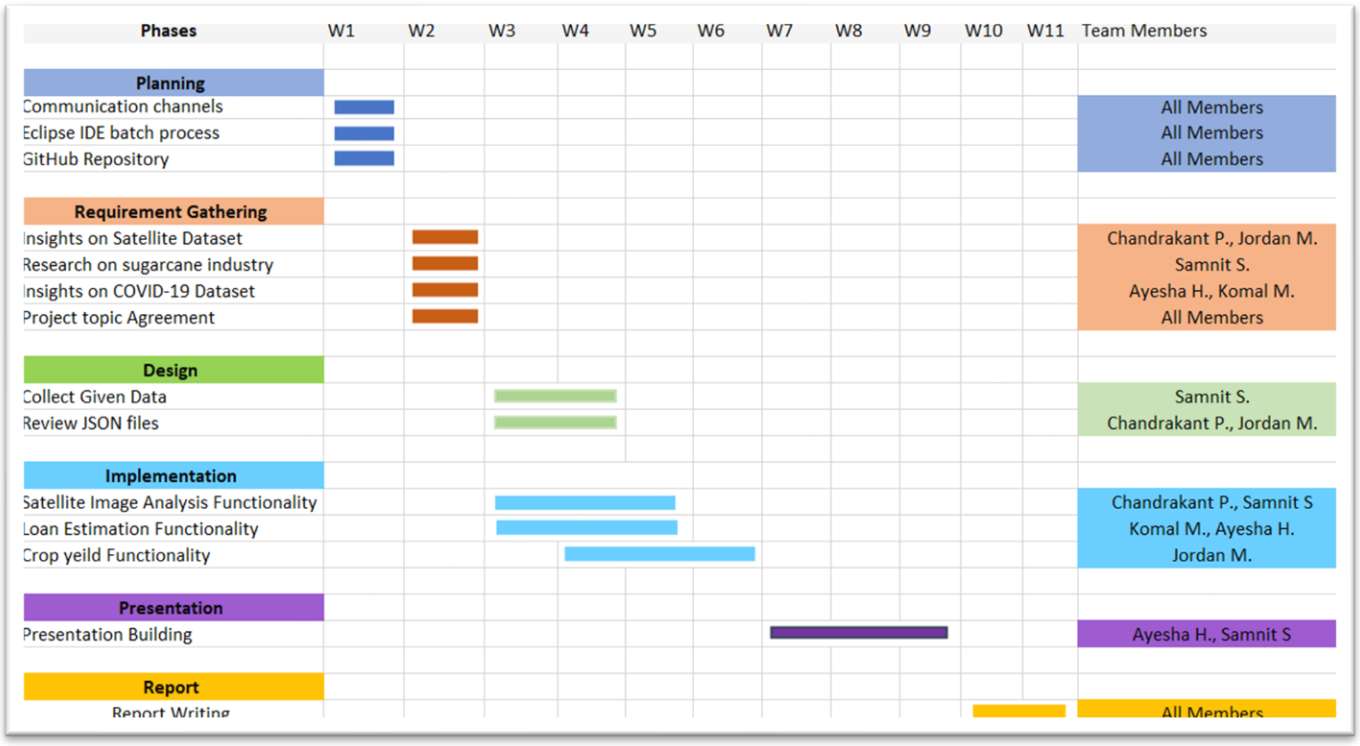

Figure 6.4: Team Contribution

Challenges faced during the project lifecycle:

1) Physical presence was not there due to COVID-19
2) JOB Timing
3) Conflicting with the Project Meeting

7) References:

[1] Australian Sugar Milling Council. 2020. *Australian Sugar Milling Council | The Peak Body For Raw Sugar Producers And Exporters*. [online] Available at: <https://asmc.com.au/?gclid=EAIaIQobChMIo43moP_B7AIVVQwrCh1YPQqOEAAYASAAEgLCFvD_BwE> [Accessed 21 October 2020].

[2] Compare the Difference Between Similar Terms. 2020. *Difference Between SDLC And Agile Methodology | Compare the Difference Between Similar Terms*. [online] Available at: <https://www.differencebetween.com/difference-between-sdlc-and-vs-agile-methodology/> [Accessed 21 October 2020].

[3] *Canegrowers.com.au*, 2020. [Online]. Available: http://www.canegrowers.com.au/icms_docs/310175_economic-contribution-of-the-sugarcane-industry-to-queensland.pdf . [Accessed: 21- Oct- 2020].

[4]"Cane Farming — Australian Cane Farms", *Australian Cane Farms*, 2020. [Online]. Available: https://www.australiancanefarms.com.au/gallery. [Accessed: 21- Oct- 2020].

[5]2020. [Online]. Available: https://www.researchgate.net/publication/325756048_A_Review_of_Bank_Loans_to_Farmers_Implications_for_Agricultural_Diversification_in_Nigeria . [Accessed: 21- Oct- 2020].

[6]"giswqs/geemap", *GitHub*, 2020. [Online]. Available: https://github.com/giswqs/geemap . [Accessed: 21- Oct- 2020].

[7]"Satellite Intelligence – Medium", *Medium*, 2020. [Online]. Available: https://medium.com/satellite-intelligence. [Accessed: 21- Oct- 2020].

[8]"Datathon Guide", *Medium*, 2020. [Online]. Available: https://medium.com/satellite-intelligence/datathon-guide-ac6539cfd623. [Accessed: 21- Oct- 2020].