**Progress Report**

**Lake Water Quality Forecasting using Remote Sensing Technology and Onsite Measurement using Machine Learning Algorithms**

**1. Machine Learning:**

The goal is to develop a machine learning model that can predict the Water Quality Index (WQI) for lake water. The dataset is being prepared to train and test the machine learning model.

* **Data Collection:** The first step in the dataset preparation was to collect the relevant data from the various resources available on internet. This involved working closely with the Indian Government data for water sources to identify the necessary data sources and extract the data.

The data sources included:

* + - * ENVIS Centre on Control of Pollution Water, Air and Noise, Hosted by Central Pollution Control Board, Sponsored by Ministry of Environment and Forests, Govt of India
      * Digital India Initiative: https://data.gov.in/
      * Karnataka State Pollution Control Board(KSPCB): <https://kspcb.karnataka.gov.in/>

Dataset for ML model involves features/attributes listed below:

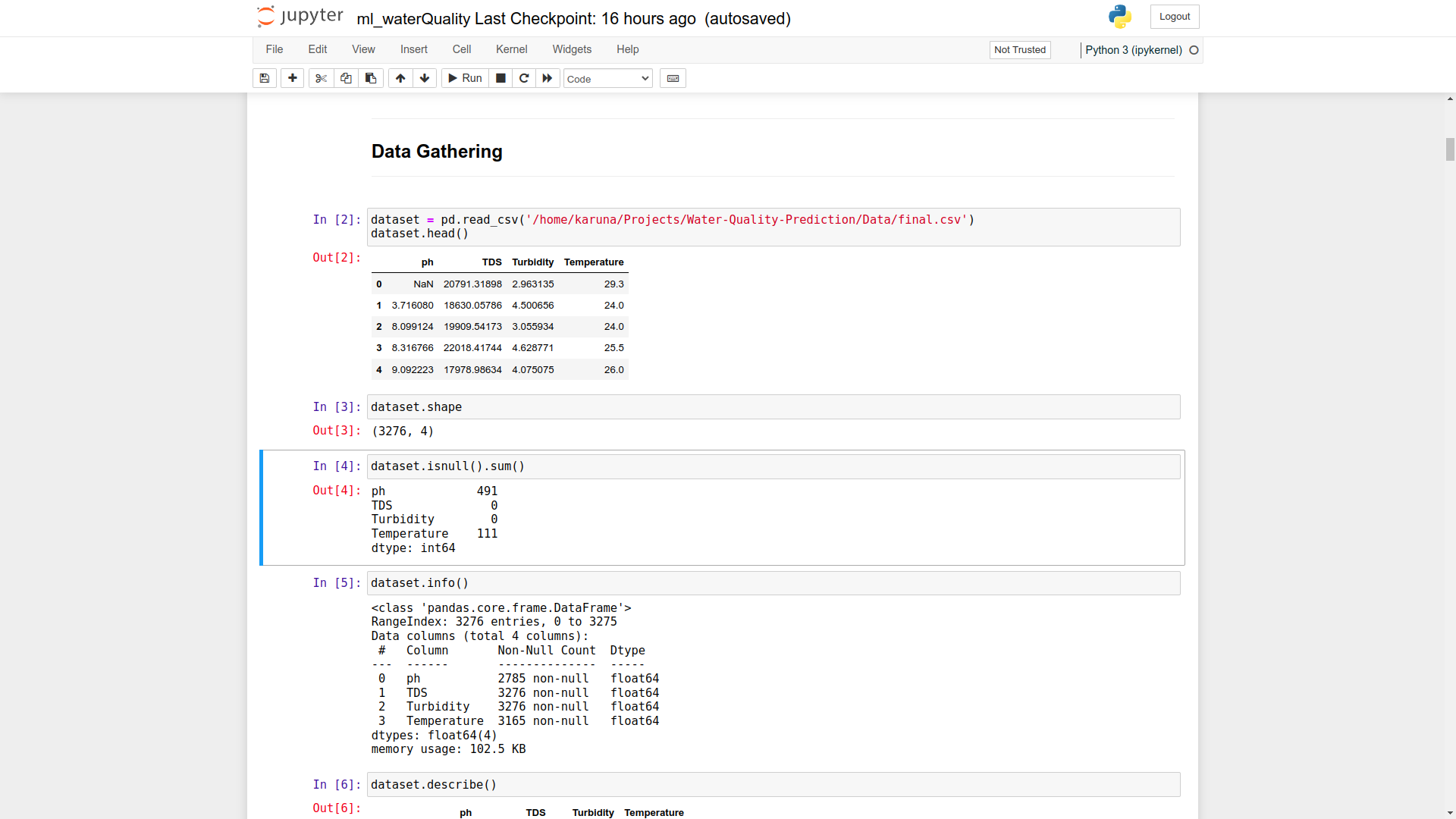
* + - * pH
      * TDS (Total Dissolved Solids)
      * Turbidity
      * Temperature

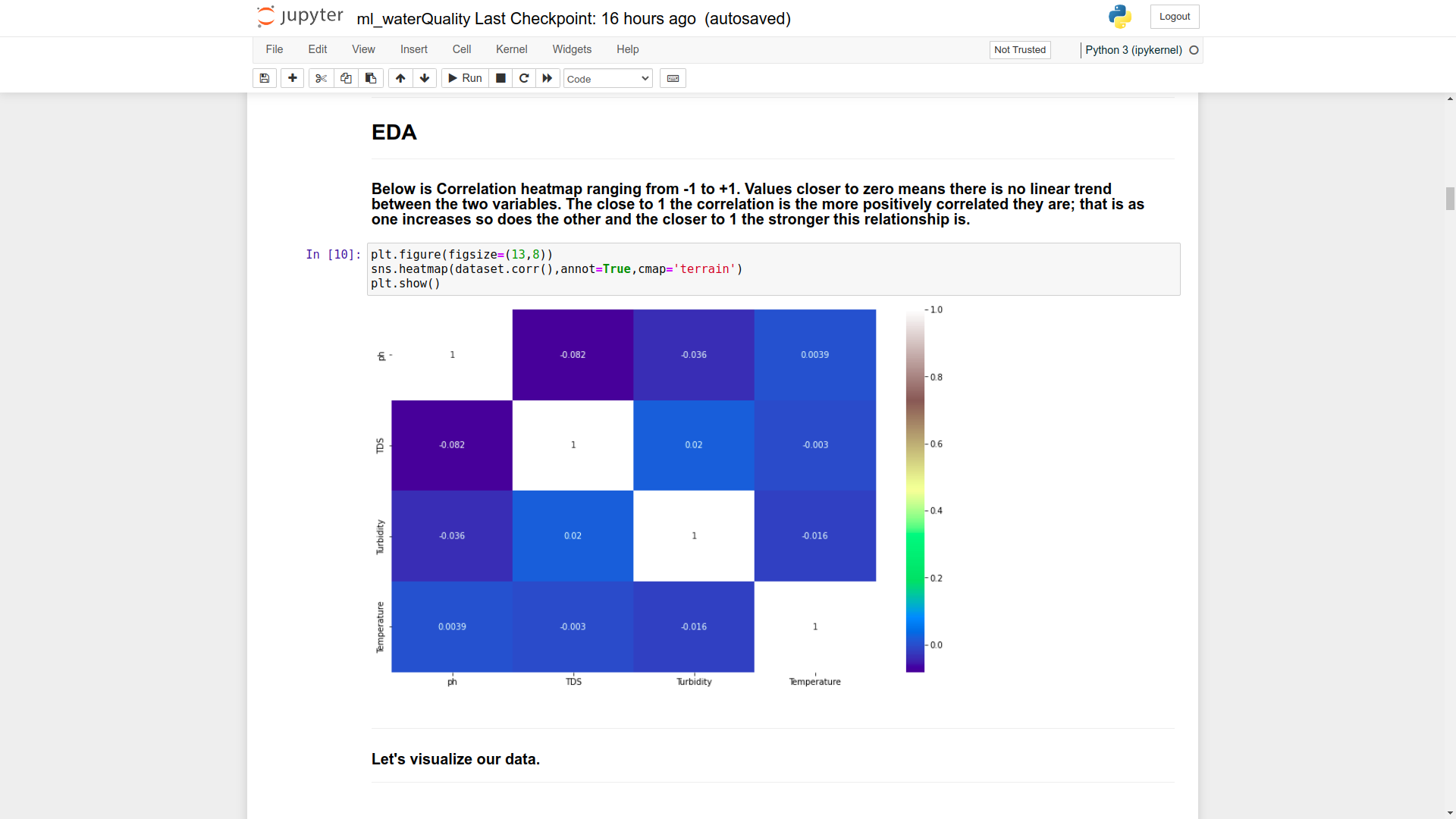
Also above features are collected from sensors onsite.

* **Data Cleaning:** After the data was collected, the next step was to clean it. This involved identifying and correcting any missing or incorrect data, removing any duplicates, and transforming the data into a usable format. This process was time-consuming but crucial for ensuring that the machine learning model is accurate and reliable.



* **Feature Engineering:** The third step was to perform feature engineering on the dataset. This involved selecting the relevant features that would be used to train the machine learning model.
  + - * Exploratory data analysis (EDA) is performed, and it reveals that the features are largely unrelated.
      * Boxplot is analysed.





* + - * **WQI** is calculated based on features:

The data which we want to predict is not already present in the data. So, before we create a model to train, we need what we want to train. So, WQI is calculated.

The WQI has been calculated by using the standards of drinking water quality recommended by the **World Health Organisation (WHO), Bureau of Indian Standards (BIS)**.

The Water Quality Index was calculated by aggregating the quality rating with the weight linearly,

WQI=∑ (qn x Wn )

where qn =Quality rating for the nth Water quality parameter,

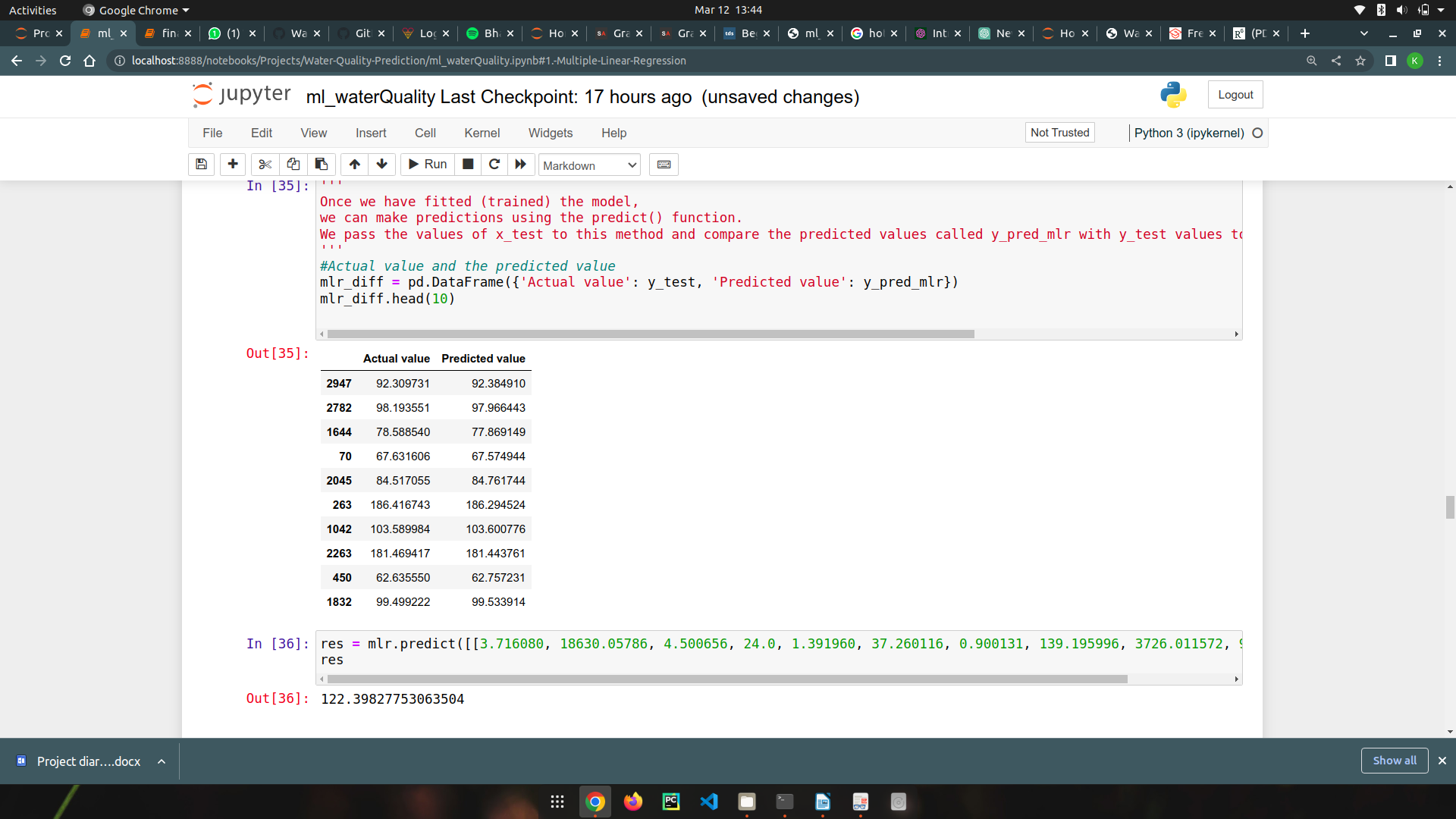
Wn = unit weight for the nth parameters.

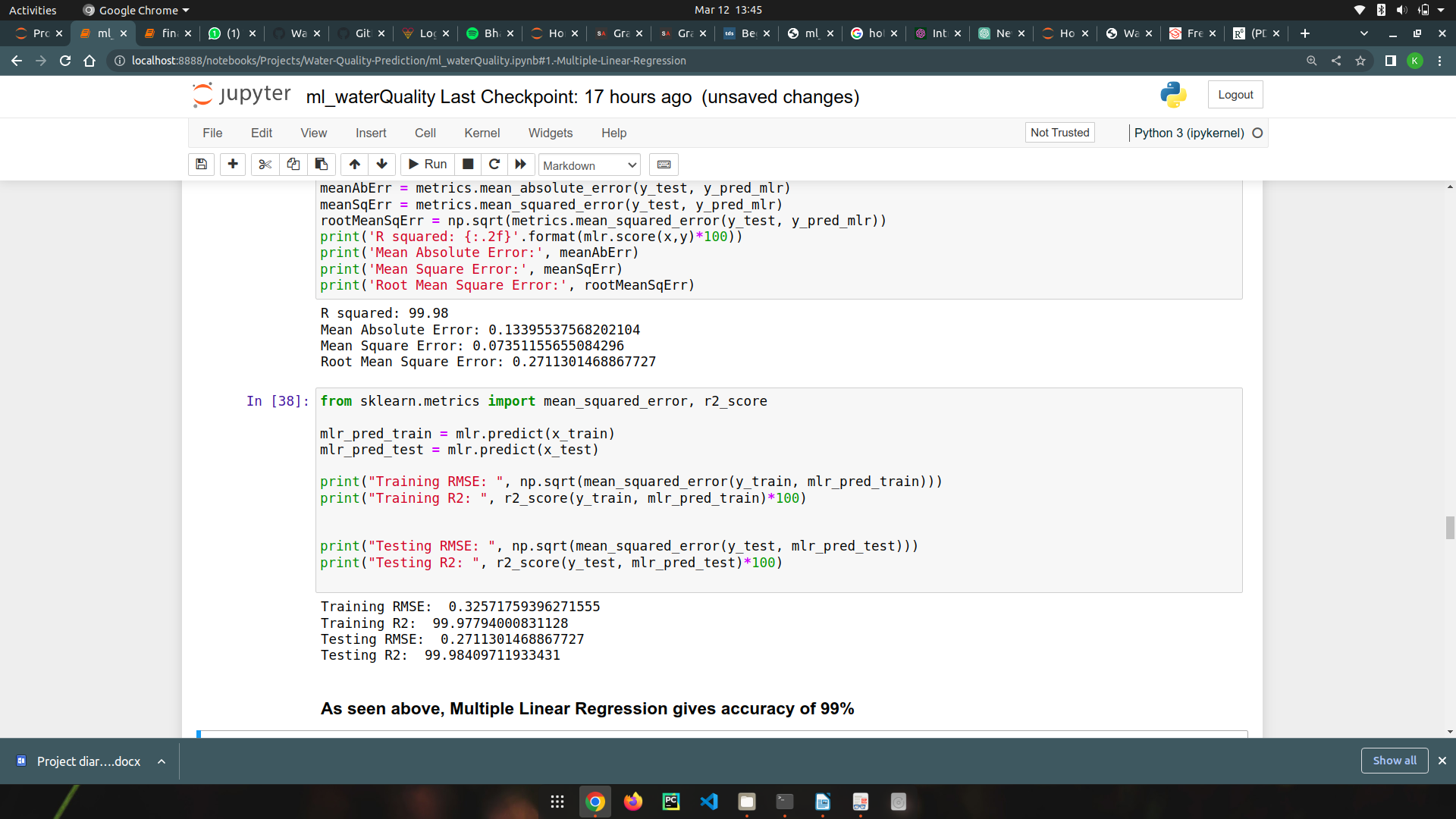
#### Weighted Arithmetic Index method is used for calculating Quality rating and unit weight for nth parameters.



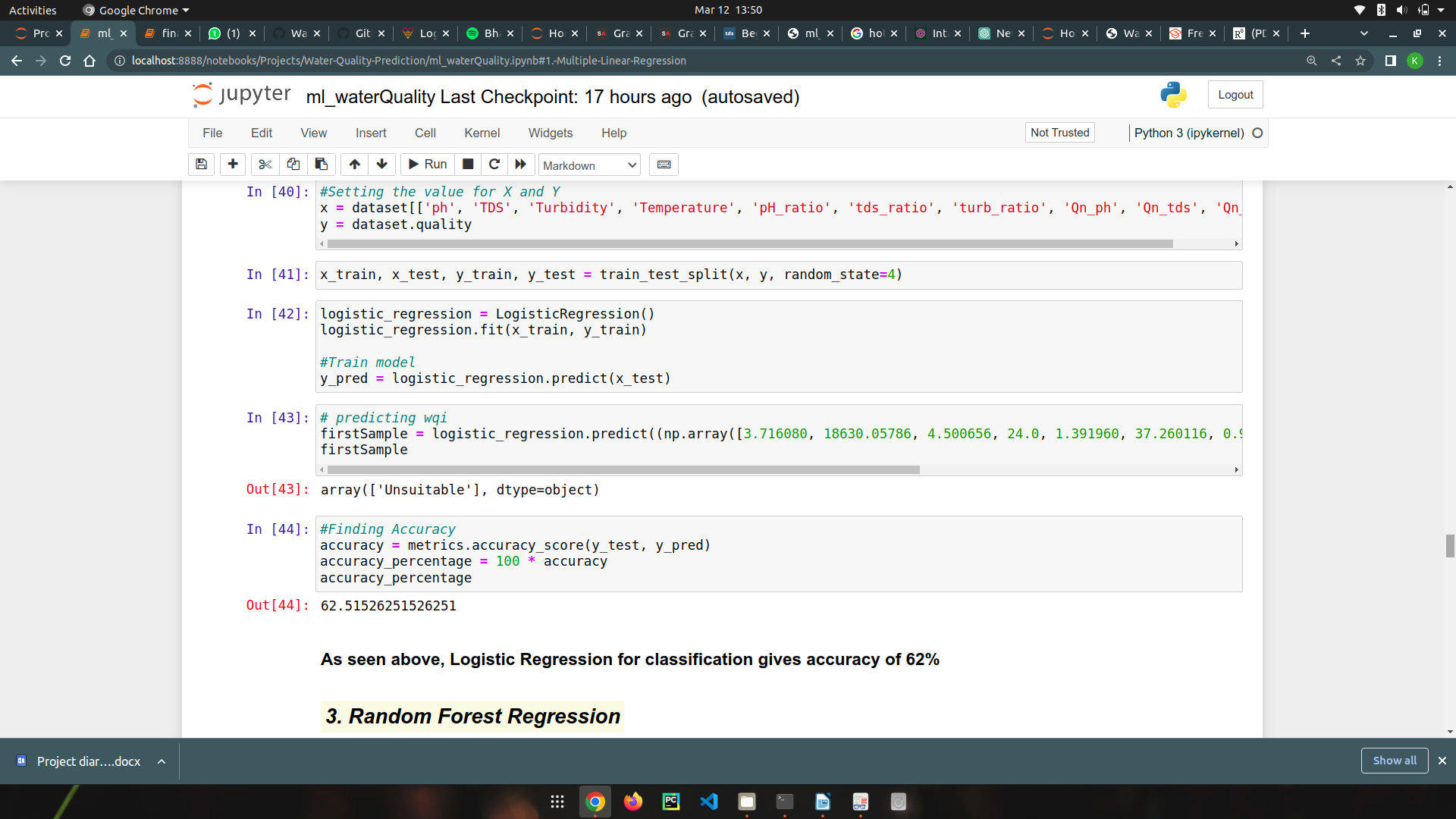
**

* **Data Splitting:** Once the dataset was cleaned and engineered, the next step was to split the data into training and testing sets. The training set was used to train the machine learning model, while the testing set was used to evaluate the model's performance. This was done to ensure that the model can generalize to new data and make accurate predictions.
* **Model Selection**- To select a suitable model for our project, we identified several potential models and used metrics such as accuracy and precision to compare their performance.
* **Model Training:**
  + - * **The first step in model training is to gather and preprocess the data. In this project, we collected a large dataset of water sample from various sources, including online repositories and user-generated content. We then preprocessed the data to remove any duplicates, outliers, and noisy samples. The final dataset consisted of over 3,000 sample, with each sample labeled according to its content.**
      * **Algorithms and Techniques / Model Definition:**
      1. **Multiple Linear Regression**
         1. **Model Evaluation: 99% Accuracy**

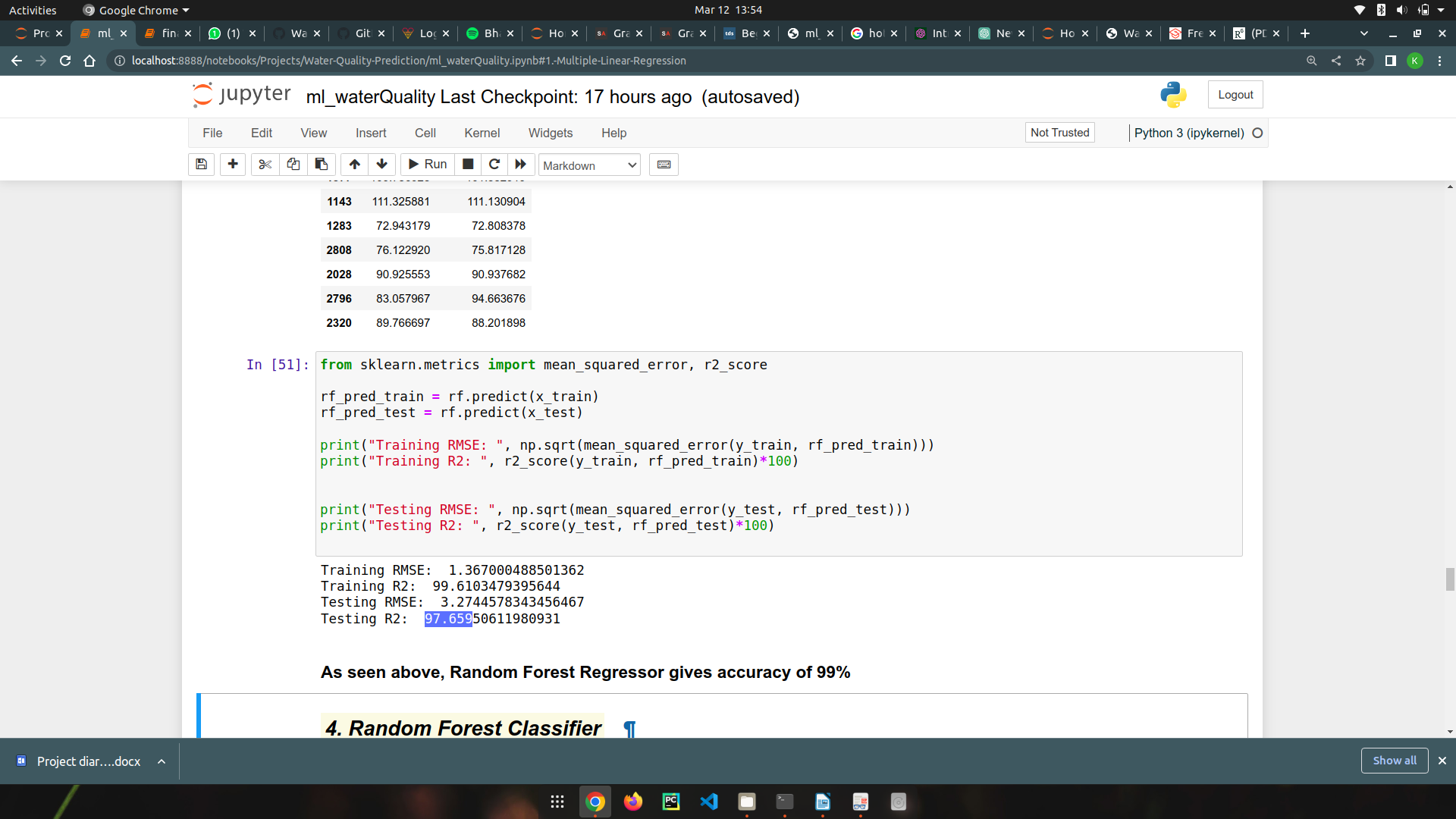




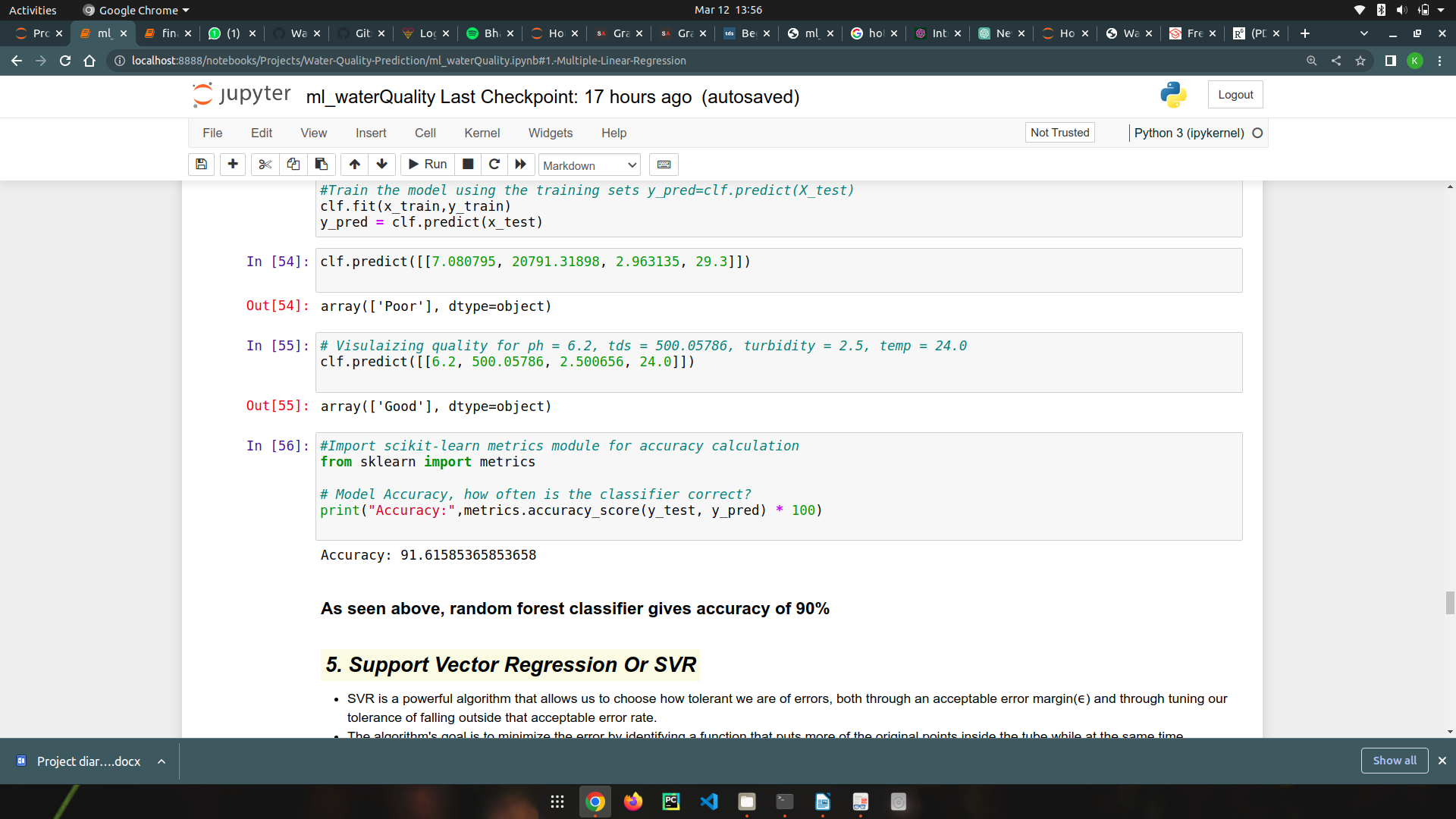
* + - 1. **Logistic Regression Model**
         1. **Model Evaluation: 62.51526 Accuracy**



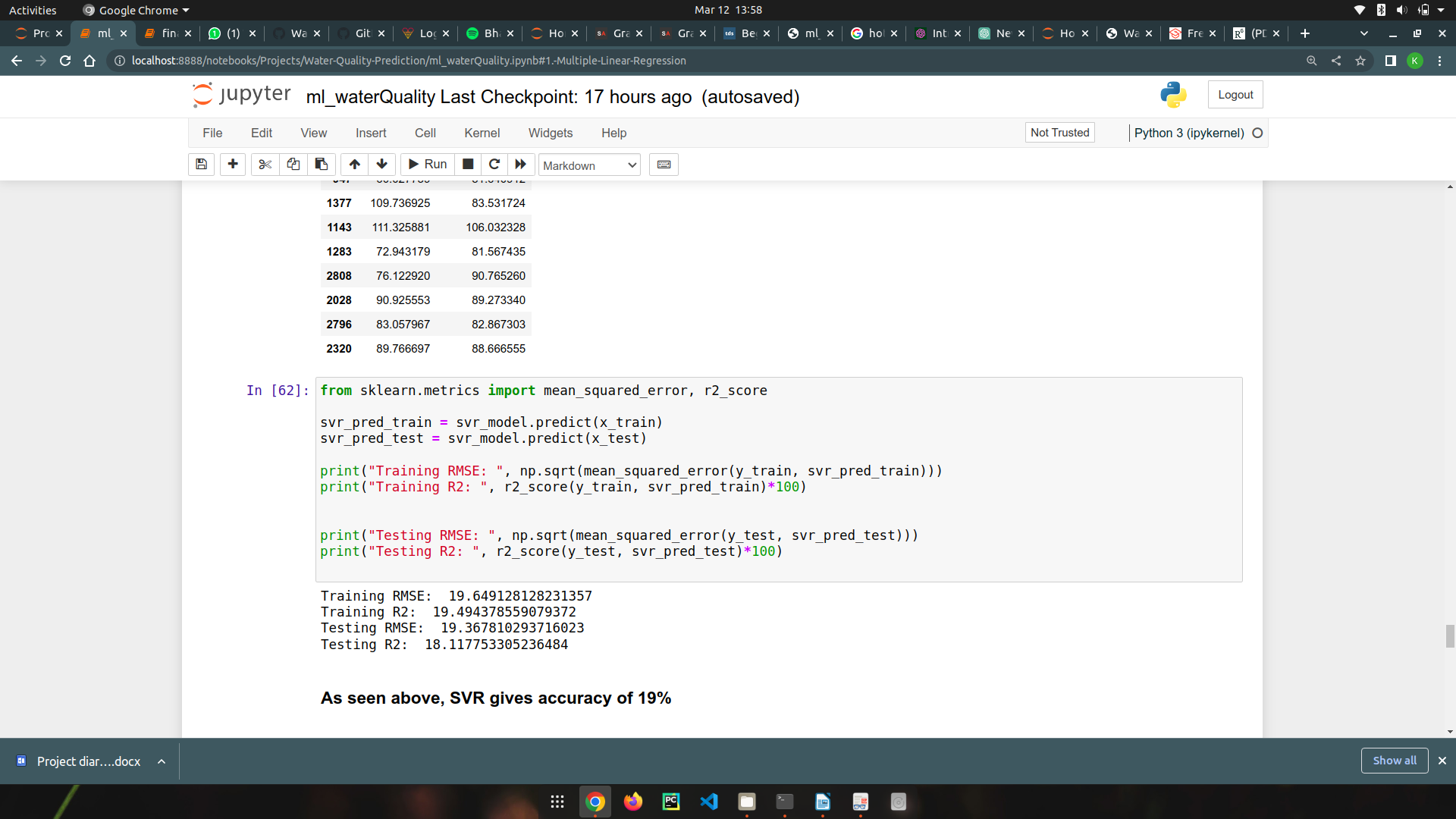
* + - 1. **Random Forest Regression**
         1. **Model Evaluation: 97.659% Accuracy**



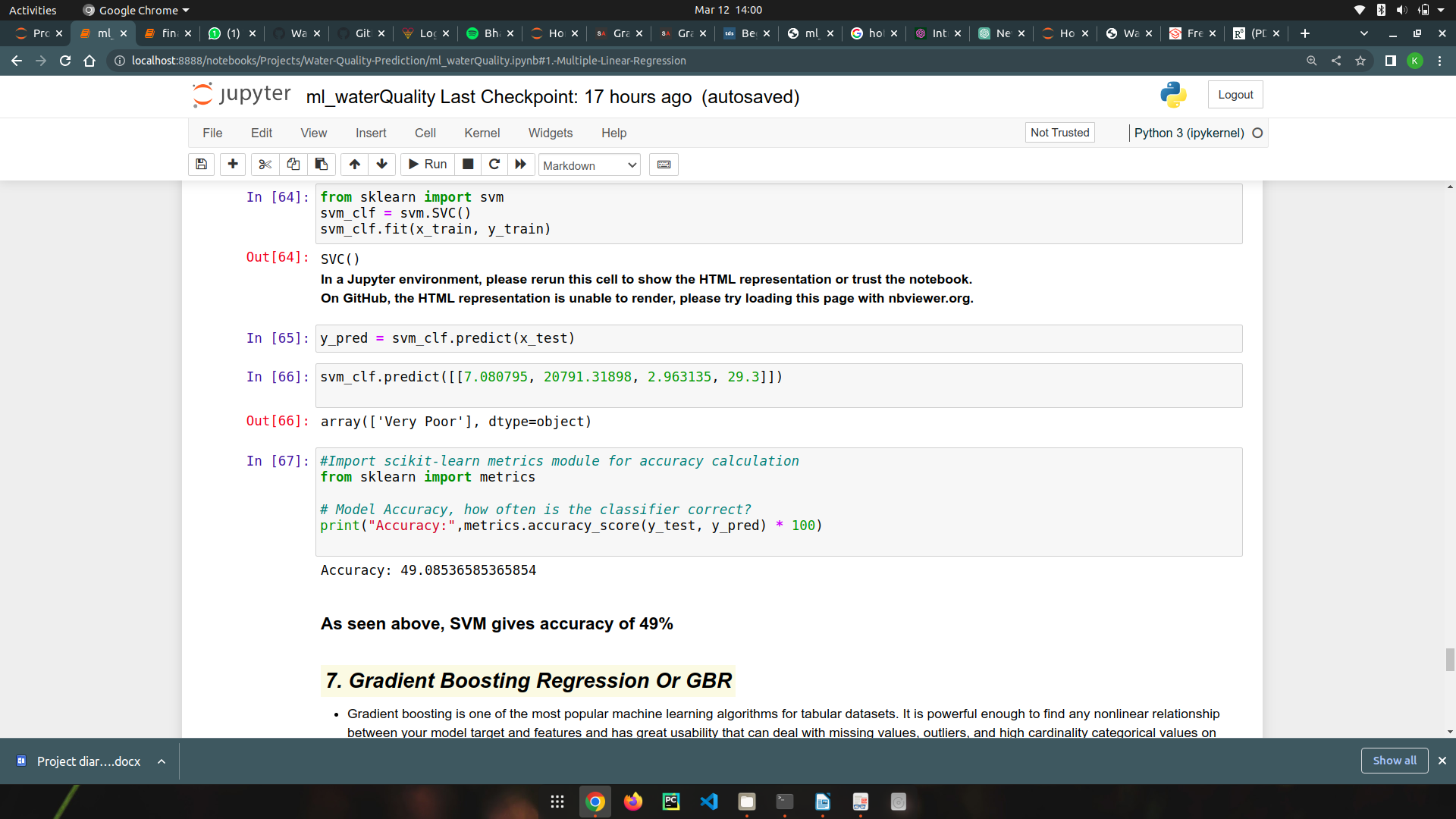
* + - 1. **Random Forest Classifier**
         1. **Model Evaluation: 91.6% Accuracy**



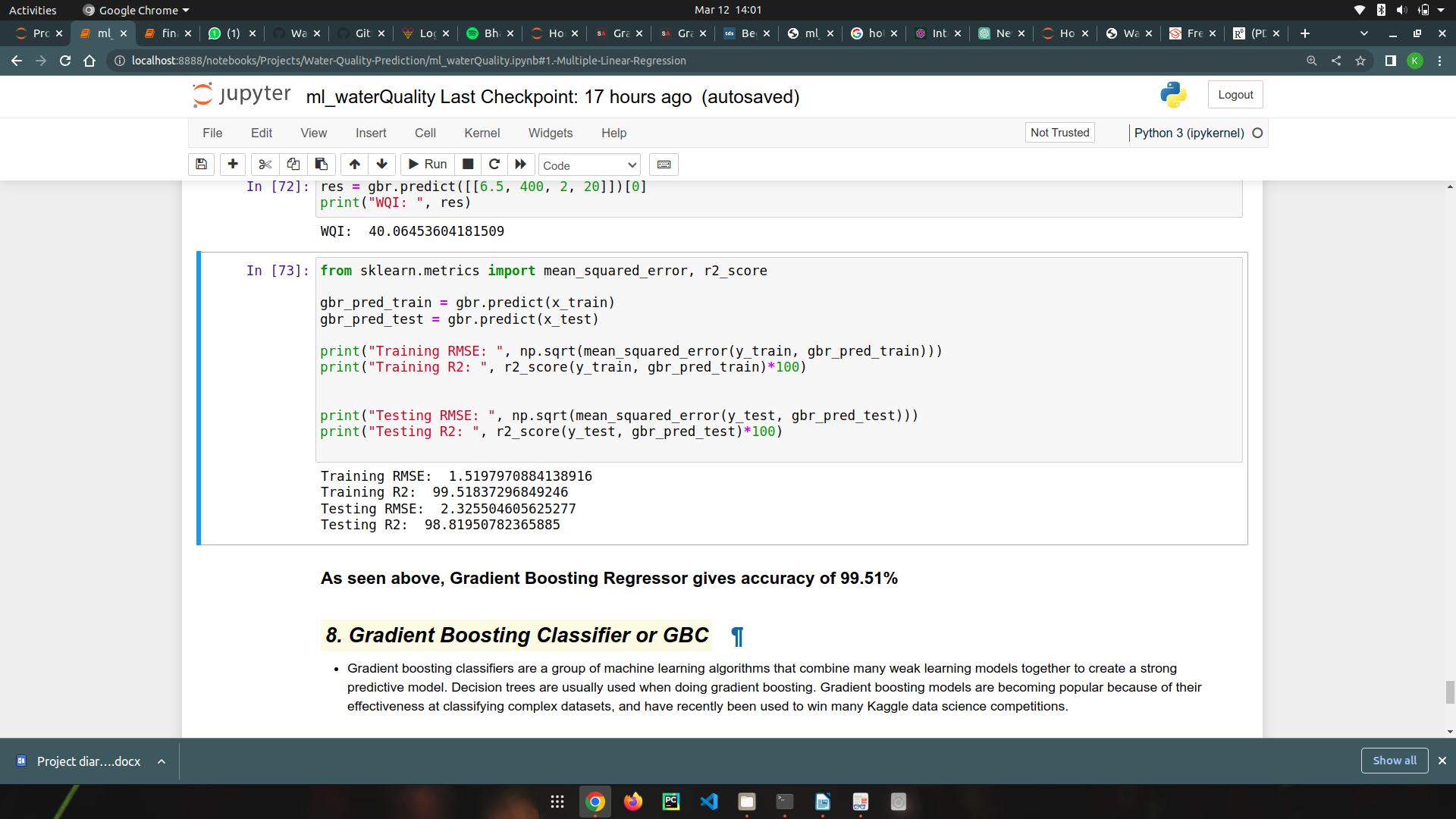
* + - 1. **Support Vector Regression (SVR)**
         1. **Model Evaluation: poor Accuracy**

**

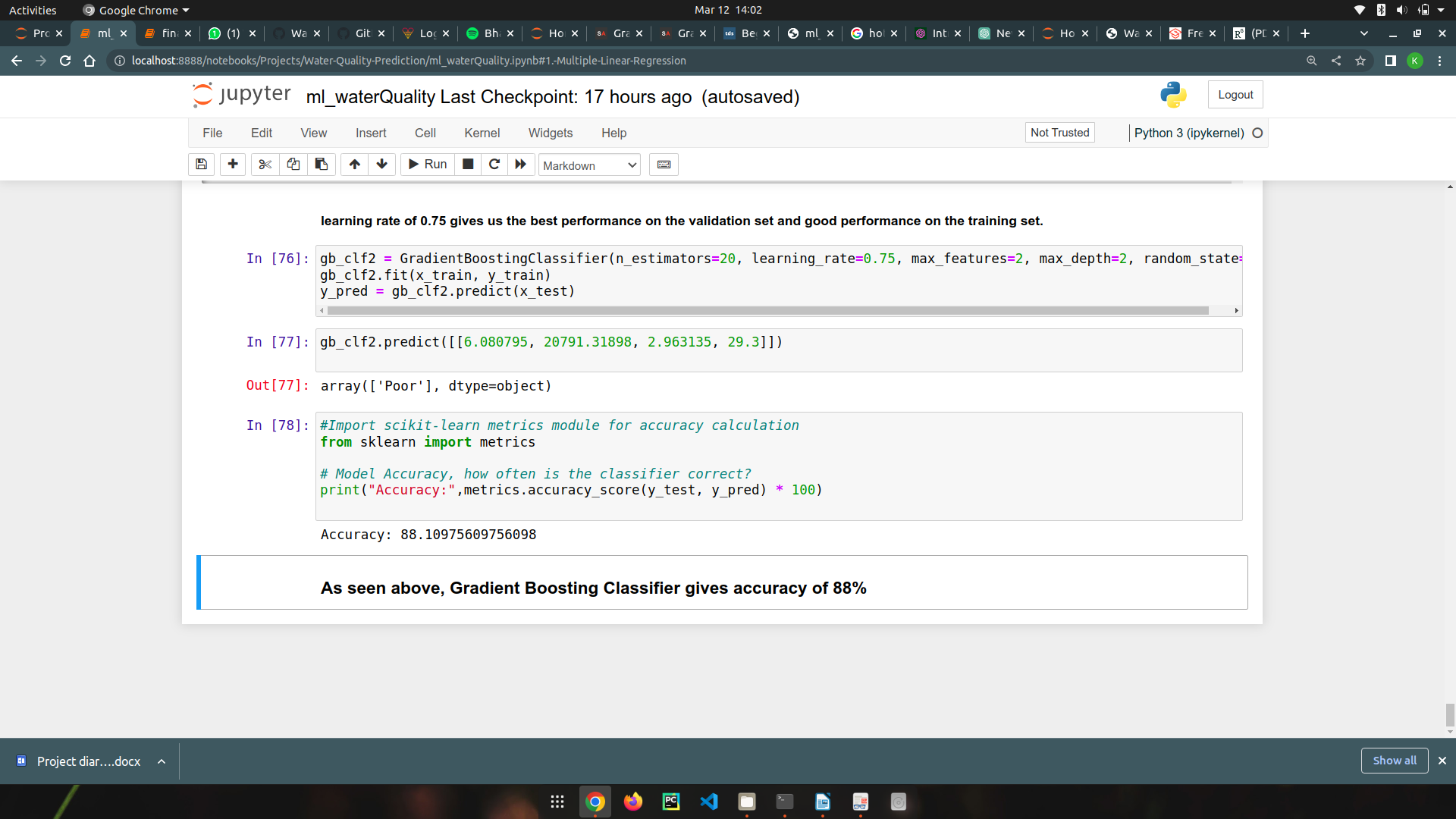
* + - 1. **Support Vector Machine (SVM)**
         1. **Model Evaluation: 50%**



* + - 1. **Gradient Boosting Regression (GBR)**
         1. **Model Evaluation: 98.819%**



* + - 1. **Gradient Boosting Classifier (GBC)**
         1. **Model Evaluation: 88.1097**

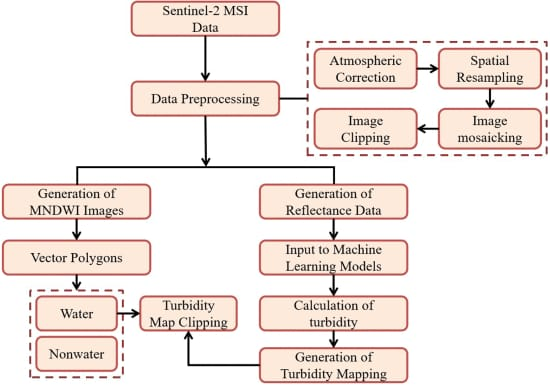
****

2.

Link : <https://github.com/Phoenixces/Water-Quality-Prediction>

**2. Remote Sensing**- The goal is to extract the Satellite Imagery of various lakes in Bangalore which provides us with the data of Turbidity and Chlorophyll. This helps us with the following reasons-

* **Large-scale coverage**: Remote sensing allows us to gather information over a large area in a short amount of time, which would be impossible to achieve through ground-based surveys.
* **Cost-effective**: Remote sensing can be a cost-effective method of collecting data as it eliminates the need for expensive on-site visits, and it is often more efficient than other methods of data collection.
* **Timely information**: Remote sensing allows us to collect information quickly, which is essential in situations where time is a critical factor, such as natural disasters or emergencies.
* **Repeated observations**: Remote sensing instruments can collect data repeatedly, allowing for the monitoring of changes over time. This feature is particularly useful for tracking environmental changes, such as land-use patterns, vegetation growth, and natural disasters.
* **Non-invasive**: Remote sensing allows us to collect information about the environment without disturbing it, which is essential in sensitive areas such as wildlife habitats, protected areas, and archaeological sites.
* **Multi-spectral data**: Remote sensing instruments can collect data in various wavelengths, allowing us to study different aspects of the environment, including vegetation health, temperature, and moisture content.



**DATA PREPROCESSING** –

Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission supporting Copernicus Land Monitoring studies, including the monitoring of vegetation, soil, and water cover, as well as observation of inland waterways and coastal areas.

**Algorithms used –**

**Normalized Difference Turbidity Index (NDTI):** This algorithm is based on the difference between the reflectance in the blue and red wavelengths, which are sensitive to the scattering and absorption of light by suspended particles. The formula for NDTI is (R865 - R469) / (R865 + R469), where R865 is the reflectance in the near-infrared wavelength (865 nm) and R469 is the reflectance in the blue wavelength (469 nm).

**Normalized Difference Chlorophyll Index (NDCI):** This is an algorithm that uses remote sensing data to estimate the concentration of chlorophyll in water bodies. Chlorophyll is a pigment found in plants and algae, and its concentration is a key indicator of water quality and productivity. The NDCI algorithm is based on the difference between the reflectance in the red and near-infrared wavelengths, which are sensitive to the absorption of light by chlorophyll.

The formula for NDCI is as follows:

NDCI = (R780 - R717) / (R780 + R717)

where R780 is the reflectance in the near-infrared wavelength (780 nm), and R717 is the reflectance in the red wavelength (717 nm).

Chlorophyll-a absorbs light in the blue and red parts of the spectrum and reflects light in the green part of the spectrum. The NDCI algorithm takes advantage of the fact that chlorophyll-a also absorbs light in the near-infrared part of the spectrum, which can be measured using remote sensing data.

**Normalized Difference Water Index (NDWI)**: This is an algorithm that uses remote sensing data to estimate the amount of water content in vegetation and soil. The algorithm is based on the difference between the reflectance in the near-infrared (NIR) and shortwave infrared (SWIR) wavelengths, which are sensitive to the amount of water in vegetation and soil.

The formula for NDWI is as follows:

NDWI = (NIR - SWIR) / (NIR + SWIR)

where NIR is the reflectance in the near-infrared wavelength (usually around 800-900 nm) and SWIR is the reflectance in the shortwave infrared wavelength (usually around 1600-1700 nm).

The NDWI algorithm is commonly used to monitor vegetation health and water content, as well as to map surface water bodies. In vegetation, water content is a key indicator of plant health and productivity, as it affects photosynthesis, growth, and stress tolerance. In soil, water content affects nutrient availability, infiltration, and erosion. The NDWI algorithm takes advantage of the fact that water absorbs light in the SWIR part of the spectrum, while vegetation absorbs light in the NIR part of the spectrum. Therefore, the difference between the reflectance in these two bands can be used to estimate the amount of water in vegetation and soil.

**Water Quality Index (WQI):** This algorithm is based on the spectral properties of water and uses a combination of visible and near-infrared wavelengths to estimate turbidity. The formula for WQI is WQI = 2.42 - (0.52 × ln(R531/R551)), where R531 is the reflectance in the green wavelength (531 nm) and R551 is the reflectance in the yellow wavelength (551 nm).

**Stage 1 - Identification of river Kempambudhi Lake in Karnataka via Google Earth Engine Platform.**

Map

Description automatically generated

**Stage 2 – Extraction of the shape file of river Kempambudhi Lake in Karnataka via Google Earth Engine Platform.**

Map

Description automatically generated

**Stage 3 – Identification of the water body using the Normalized Difference Water Index (NDWI) algorithm**

Map

Description automatically generated

**Stage 4 – Sentinel 2 MSI functions and algorithms implemented through the cloud pixel percentage of the lake and based on the NDCI (Normalized difference in chlorophyll index) and Normalized Difference Turbidity Index (NDTI) we determine the turbid state of the lake by determining it based on the color scheme.**

Map

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Stage 5 – Importing and connecting images to ArcGIS for further map clipping images processing process.**

**3. Website Development:**

I am pleased to report that we have completed the frontend development of our website. Our team of frontend developers worked tirelessly to ensure that the website meets the desired specifications and is user-friendly.

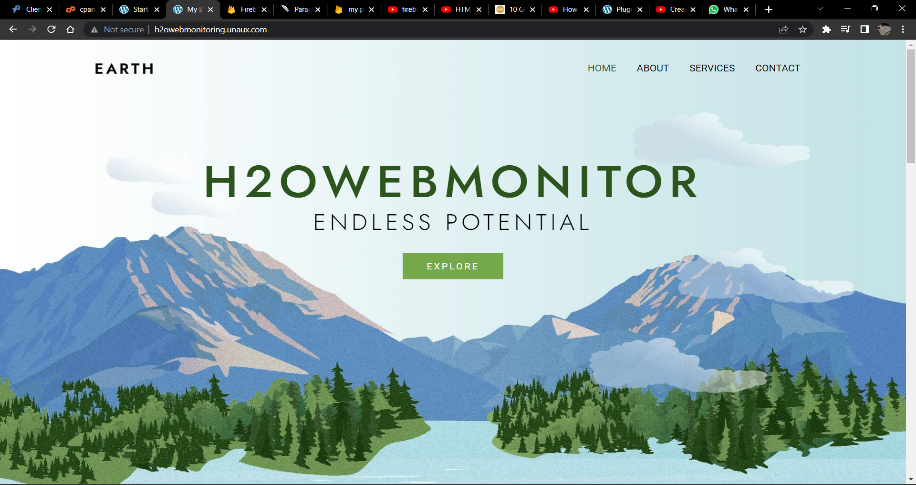
We have implemented a modern, responsive design that is optimized for both desktop and mobile devices. The website's layout and navigation have been streamlined to ensure that users can easily find the information they are looking for. We have also made sure that the website is visually appealing with the use of high-quality images and graphics.

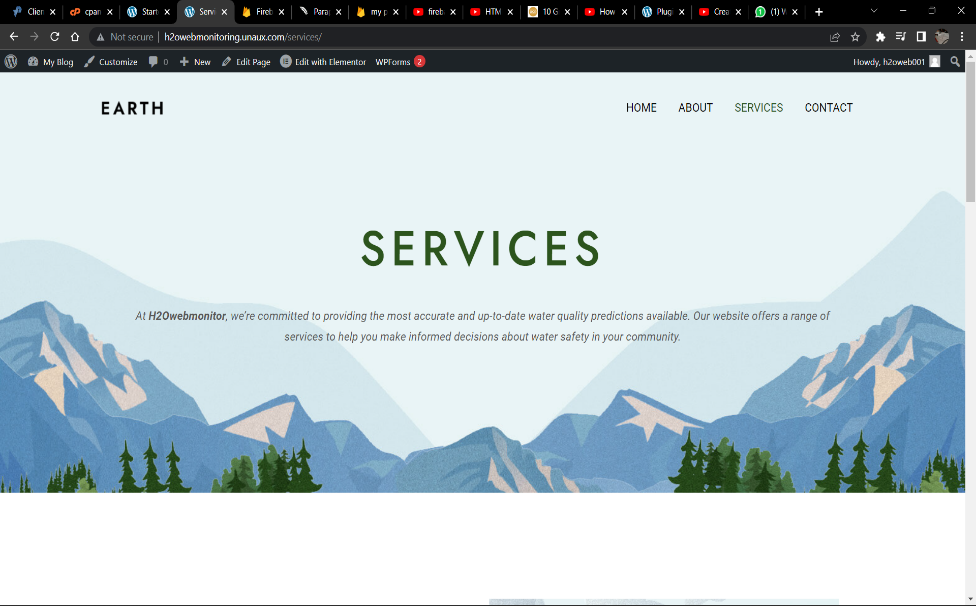
During the development process, we have tested the website on multiple devices and browsers to ensure that it is fully functional and performs well under different conditions.

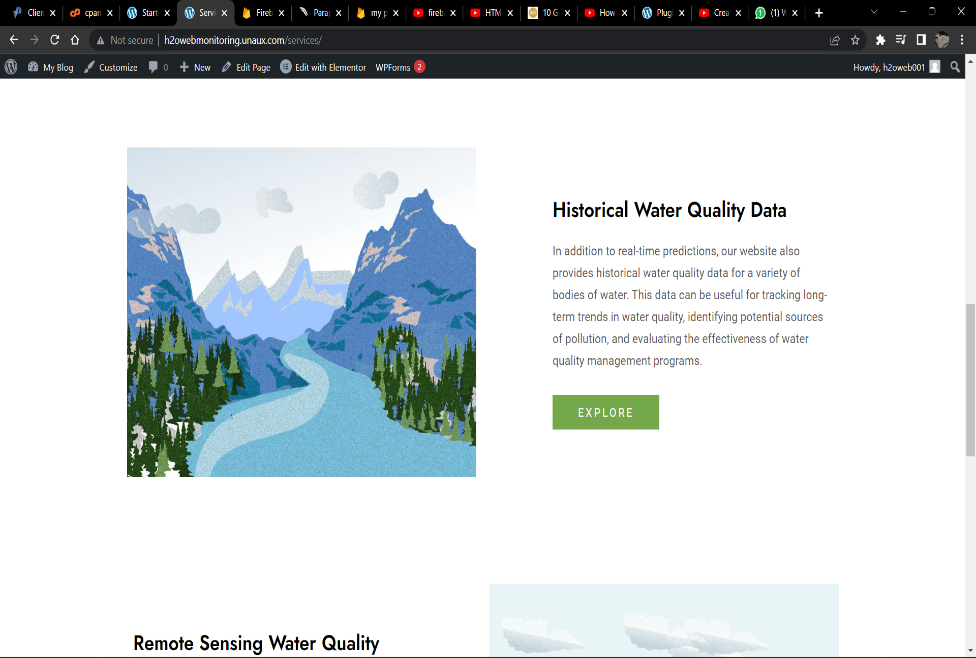
Moving forward, we will be dealing with the backend development of the website using Firebase software. While we expect this to take some time, we are confident that we can deliver a high-quality result that meets our client's needs through this project.

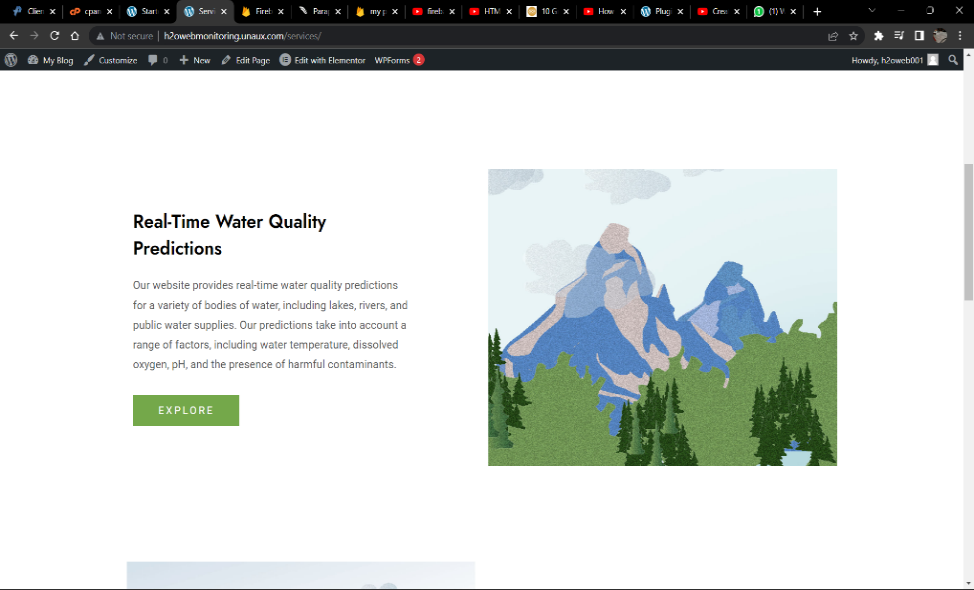
Thank you for entrusting us with this project. We look forward to continuing our work with you and delivering a successful outcome.

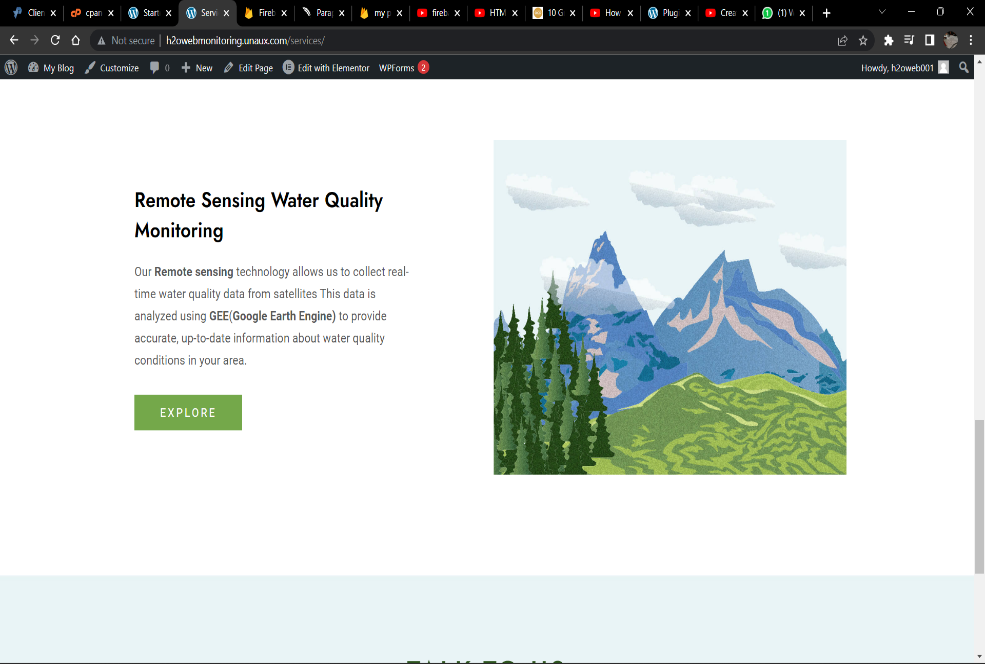
Few images of our Web-site:











Link : <http://h2owebmonitoring.unaux.com/>

(The hosting as of now is free hosting)

4. Image Processing:

5. Smart Buoy for onsite Measurement: