University of Tunis El Manar Faculty of Sciences of Tunis Department of Geology

Final Year Project geomatics license

Option: GAST

Created By: khiari mohamed

Topic:

Hybrid Quantum-Classical Neural Network for resources exploration

- FST supervisor: Assistant professor MR Tarek sboui
- Company supervisor : Associate professor Mr. Moncef Ben Smida
- HOLD
- HOLD ....

University year 2023-2024

This project was made in collaboration with C.I.P.E.M

1. Acknowledgements:

First of all, I want to thank God because He taught me that a debt of gratitude must be repaid. Through every trial and triumph, I've come to realize that with God by my side, I'm never alone.

I'm thankful for everything—the highs and lows, the comebacks and setbacks that made me who I am today.

I eagerly anticipate the presence of my family. To my dad, who taught me how to be resilient, thank you for making countless sacrifices along the way.

To the most important person in the whole world, my mother, a wonder woman, whose boundless love and steadfast prayers have been my guiding light throughout There are 8 billion people on this planet, and you are the perfect mom for me.

I wish to offer a profound acknowledgement to Mr. Moncef Ben Smida. His generosity in offering me the internship opportunity at CIPEM has been truly appreciated. returning the favor for his guidance and mentorship I'm thanking him today for making me developing insights and skills that have significantly enriched my understanding of the industry. I am thankful for the confidence he put in me during my tenure at CIPEM.

My heartfelt gratitude extends to Mr. Tarek Sboui, my supervisor. I am deeply grateful for your enduring patience, invaluable guidance, and devoted encouragement. Your valuable lessons have not only increased my interest in the field but also fueled my desire to excel in it. Thank you for being a building block of my academic and professional journey.

Lastly, I am deeply grateful to my roommates, who have stood by me through every challenge because this honor doesn't just belong to me i wouldn't come this far If it weren't for certain significant individuals in my life, beginning with my mother, father, brother, and my other family, who I'm so glad to have

i was under the misconception that my achievements were solely my own. Nothing can be further from the truth. I've been uplifted, supported, motivated, and embraced by an extraordinary circle of companions unlike any other Rayen Dabbabi, Rayen Ben Hassen, and Mouath Haffar

being here today, brimming with gratitude, I realize that my journey is just beginning. I carry the lessons, love, and support you've all given me as precious fuel for the road ahead. As I step forward, I promise to honor your encouragement, pay it forward with kindness, and strive to make you all proud. Thank you, for being the source of empowerment.

Table of Contents

List of figures

Chapter 1: Introduction

General Context

In the evolving field of energy exploration, the search, for environmentally friendly methods of extracting hydrocarbons has brought together traditional geoscience and advanced technology.

With the demand for energy on the rise, there is a growing need to innovate hydrocarbon exploration methods. This initiative marks an effort that combines geoscience with modern artificial intelligence. Historically, discovering and describing oil and gas reservoirs has heavily relied on established models and time-consuming surveys. However, the emergence of machine learning offers an opportunity to enhance these approaches with data-driven insights and predictive analysis.

This project explores the potential of quantum machine learning (QML) in hydrocarbon exploration. Using quantum mechanics has unique features for managing datasets and potentially transforming reservoir identification processes. Relying on the capabilities of QML algorithms, we aim to surpass classical machine learning limitations and achieve unique accuracy in predicting reservoirs. but also pave the way for a future where energy extraction is optimized and environmentally responsible. I suggest that you join us on this cutting-edge exploration where technology and geoscience converge as we strive to unlock new avenues for the sustainable extraction of essential energy resources.

problem and objectives

a. **Problem**

The traditional approach to hydrocarbon exploration presents financial fortitude due to the substantial upfront investments, meticulous planning, and assembly of skilled teams needed. This process demands a significant amount of money, ranging from $8 million onshore to over $100 million offshore. To navigate this economic landscape more effectively, innovative solutions like the exploration of quantum machine learning (QML) are very promising.

By potentially optimizing drilling locations, reducing exploration time, and minimizing resource waste, a groundbreaking approach can revolutionize exploration methodologies and potentially alleviate the financial burdens associated with traditional methods.

b. **Objectives:**

Revolutionize traditional hydrocarbon exploration by integrating **quantum machine learning (QML)** to significantly reduce the substantial financial investment and meticulous planning associated with the process.

Leverage the unique capabilities of QML to achieve **unprecedented accuracy** in reservoir predictions, **optimize resource extraction processes**, and **minimize environmental impact**.

Develop a novel, efficient, and cost-effective approach to hydrocarbon exploration that promotes **environmental sustainability**, thereby contributing to the **advancement of the energy industry** and addressing the limitations of traditional methods.

Chapter 2: LITERATURE REVIEW

Introduction

Finding hidden oil and natural gas resources is a major challenge. Traditional exploration methods can be very slow, taking years to get results. This process not only slows down new discoveries but also often overlooks the potential of natural gas reserves. We propose a new approach to make finding oil and gas easier by combining quantum mechanics with regular machine learning. This new hybrid strategy could change how we explore hydrocarbons by:

Accelerating the Exploration Process:

Quantum machine learning can work with difficult data quickly. It unravels complex datasets linked to oil and gas exploration. Old machine learning methods battle with intricate geological data. However, quantum algorithms like the Variational Quantum Eigen Solver (VQE) analyze this data much faster. As a result, possible locations for oil and gas reservoirs can be identified rapidly.

Expanding the Hydrocarbon Scope:

Our plan goes beyond typical ways. We look for oil and natural gas together when exploring underground. This gives us a fuller picture of what's beneath the surface.

Enhancing Predictive Accuracy:

The fusion of QML and ML holds the promise of unlocking a new level of predictive accuracy. Our enhanced technology allows for pinpointing probable mineral reservoirs more precisely.

The Power Beneath Our Feet: Geology and the Exploration of Shale Oil and Gas

Our journey begins with the fascinating world of geology, the key to unlocking the secrets hidden beneath our feet, from valuable resources to beautiful minerals. Starting with the historical significance of hydrocarbons and contrasting the properties and uses of oil and gas, covering these concepts will pave the way for examining the challenges faced by traditional methods. Next, we will delve into how AI can revolutionize this industry by overcoming long-standing limitations and unlocking new possibilities for the future.

1. geology :

Geology is the study of the Earth's structure, composition, and processes. using a variety of techniques to locate where sizable quantities of hydrocarbon exist under the surface, primarily employing seismic imaging, which can also be called an ultrasound, which involves sending sound waves into the earth and recording the echoes that bounce back. Based on these records, geologists can create a prototype of the rock layers, leading them to determine the depth and thickness of the shale formations and their size as well. Year after year, we've seen the demand for natural resources such as oil and gas progressively increase, but onshore "conventional" plays have struggled to keep up. Big firms needed a solution to continue the demand, and shifting to unconventional methods was an option to grow production volumes and secure their country's energy future. Sedimentary rocks like shale can be full of organic matter, containing remnants of plants and algal material preserved inside of them. These rocks are buried deeper beneath the earth's surface, where the presence of the best factors like pressure and temperature increases over time, backing hydrocarbons like oil and gas. After identifying the targeted areas, geologists collect rock samples during drilling. These samples provide valuable information on the composition and characteristics of the shale, like porosity and permeability.

They can also run tests on the fossils in the rock to help identify the age of the shale, leading them to a better understanding of whether it's from the from the Cenozoic or any other geologic past in which the rock was formed, the environment, categorization, and organic matter composition. Based on that, they can determine the potential quantity and quality of the matter (gas and oil) that could have been generated over time. After gathering all the data, computer

models simulate the movement of the oil. Energy companies use this information to design effective production techniques., such as hydraulic fracturing

figure1: Graphic representation of typical Avis in a general-purpose geologic map that can be used to identify geologic hazards, locate natural resources

1. **gas vs oil:**

We've talked about hydrocarbons so far, which raises the question: What's the difference between oil and gas?

biologically speaking, they both come from prehistoric organisms that decomposed underground, as we talked about earlier, but the main difference in the composition and physical state is that crude oil is a complex format of hydrocarbon compound; along with hydrogen and carbon atoms, there is also a small amount of sulfur, nitrogen, and oxygen. Natural gas, on the other hand, is a simple chemical compound of methane (CH_4); it may also have some propane. Also, crude oil is a liquid at standard atmospheric conditions, while natural gas is a gaseous state hydrocarbon lighter than air. Extraction methods can also be different for the oil. After drilling, additional techniques can be applied, like hydraulic fracturing and horizontal drilling, while for the gas, instead of releasing oil, it releases gas.

1. **A Century of Progress:**

The natural gas and oil industry has come a long way since its early days in the mid-1800s. From the very first oil well drilled in Pennsylvania in 1859 to the advanced drilling techniques used today, this industry has grown rapidly and significantly impacted the world. In the beginning, oil was mainly used for lighting and lubrication. However, as the world became more industrialized, oil became essential in producing a wide range of products, from plastics to gasoline. The discovery of natural gas, a cleaner-burning fuel, in the early 1900s further increased the importance of this industry. But environmental concerns were a conundrum along with the fluctuating prices, and despite all that, it made incredible advancements. For example, hydraulic fracturing, or “fracking,” has revolutionized the industry by allowing companies to extract natural gas and oil from previously inaccessible sources. Another major development has been the rise of renewable energy sources, such as wind and solar power.

The breakthrough wasn't as big as the oil but their increasing popularity is a testament to the changing demands of consumers and governments alike. Today, the natural gas and oil industry continues to innovate and adapt to new challenges. From using artificial intelligence to improve exploration and production processes to investing in carbon capture and storage technology, companies are working to reduce their environmental impact and ensure a sustainable future.

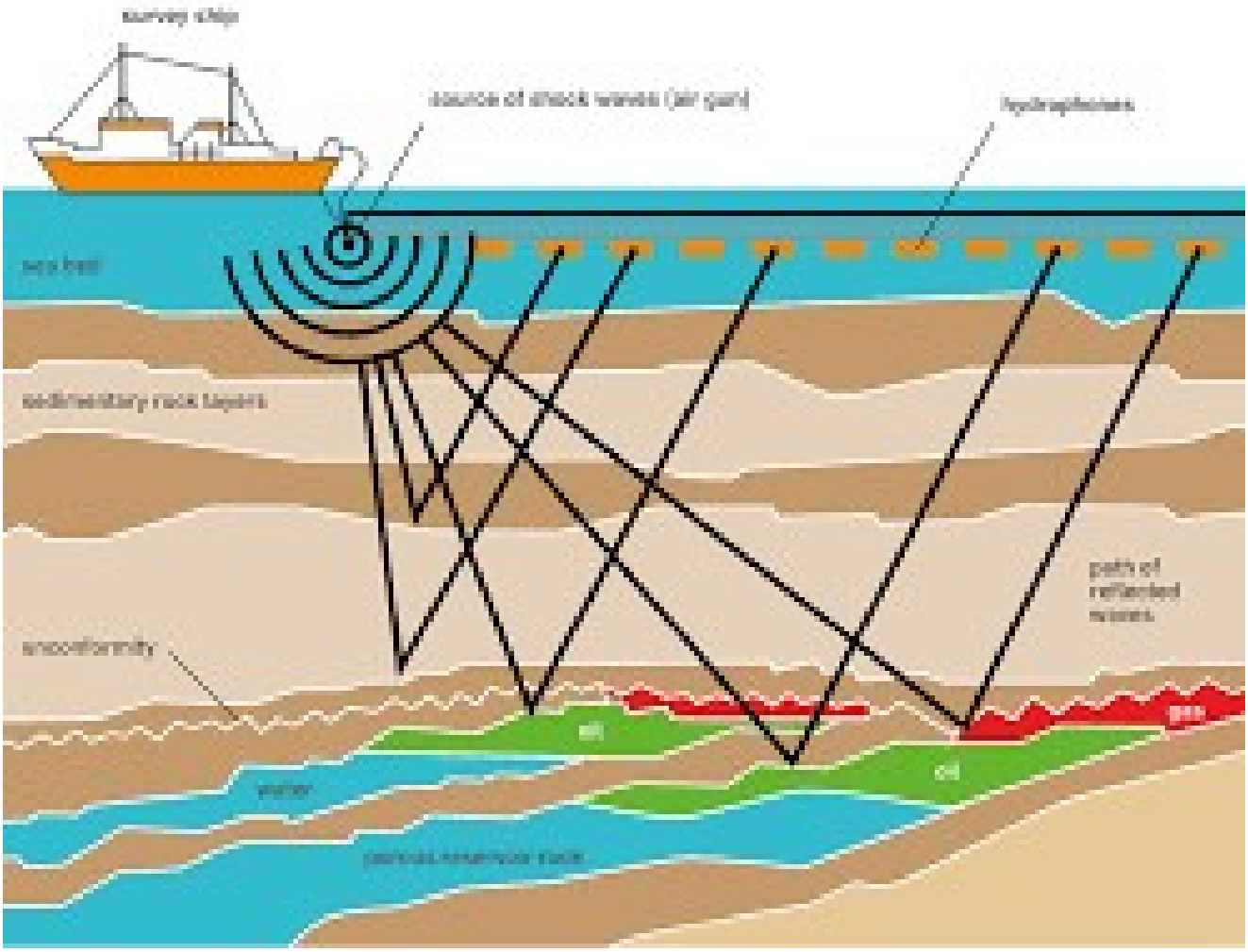


figure2: Diagram displaying the shape of numerous specific forms of oil and fuel line traps

1. Challenges and Limitations of Traditional Exploration Methods:

While traditional exploration methods have played a major role in discovering and extracting hydrocarbons for centuries, they face several significant challenges and limitations:

Difficulties in Identifying Potential Sources: Traditional methods often struggle to access resources buried deep underground or located in environmentally sensitive areas. Additionally, limitations in data acquisition and analysis can lead to inaccurate assessments of potential reserve size and quality. High Costs and Time Consumption: Activities like drilling and seismic surveys can be incredibly expensive and time-consuming; the cost alone is way too high, and the exploration process itself can take years to complete. not forgetting that the lack of reach and depth penetration capabilities needed to access certain formations effectively and interpret the data collected through these methods can be complex and prone to errors. These challenges contribute to inefficiencies and increased costs in hydrocarbon exploration, paving the way for advancements like artificial intelligence to offer transformative solutions in the future.

1. **artificial intelligence : New Solutions on the Horizon :**

Artificial intelligence (AI) is a game changer by offering a powerful set of tools to overcome the limitations of traditional methods. a subset of AI that enables algorithms to learn from vast datasets without explicit programming. By analyzing enormous volumes of geological, seismic, and production data, ML models can identify subtle patterns and correlations that would be impossible for humans to detect. This capability allows AI to predict the location of hydrocarbon reservoirs with greater accuracy, optimize drilling and extraction techniques, and reduce exploration costs.

A 2.0 form of machine learning called quantum machine learning (QML) is showing potential for further investigation. the principles of quantum mechanics, specifically the behavior of qubits, which can exist in multiple states simultaneously, unlike the bits in traditional computers. This unique property allows QML algorithms to explore a vast number of possibilities simultaneously and calculate the outcomes three times faster than average, potentially leading to breakthroughs in data analysis and problem-solving. While still in its early stages, QML has the potential to further revolutionize exploration by tackling complex geological challenges with unprecedented power and efficiency.

Chapter 3: Bridging the Gap from Remote Sensing to AI-powered Hybrid QML in Geomatics

Introduction:

This chapter connects the dots between technologies, such as remote sensing features, within the field of geomatics.

2.1 Unveiling the Earth's Secrets: Remote Sensing:

Remote sensing is a powerful tool, providing valuable insights into Earth's surface and subsurface characteristics without physical contact. This technology utilizes various platforms to collect data from a distance, helping us "see" the Earth in ways traditional exploration methods cannot.

2.2 Satellite Imagery: A Window to the Earth's Surface

Satellite imagery is a transformative technology that reshapes our perception and understanding of our planet. These images, captured by satellites, offer invaluable insights into Earth's features, anomalies, and patterns, as well as

responses to various issues.

Satellite Imagery Definition:

Literal to its name, Satellite Imagery refers to images captured by satellites, presenting a digital visual representation of the Earth's surface through cameras or sensors mounted on satellites orbiting the Earth.

Active and Passive Satellites:

Satellites are categorized as active or passive. Active Satellites use remote sensors to detect reflected responses from objects irradiated by artificially generated energy sources. In contrast, Passive Satellites use sensors to detect reflected or emitted electromagnetic radiation from natural sources, such as the sun, magnetism, or geothermal activity.

Types of Satellite Imagerie :

1. *Visible Satellite Imagery :*

Captures images using satellites to detect and record visible light wavelengths, providing a visual representation of the Earth's surface and cloud cover. Available only during the day.

2.Infrared (IR) Satellite Imagery:

Captures infrared radiation emitted or reflected by objects on the Earth's surface, highlighting temperature variations. It is valuable for weather forecasting, temperature analysis, and detecting heat signatures in different environments. Available day and night.

3.Water Vapor Satellite Imagery:

Designed to detect the concentration and movement of water vapor in the Earth's atmosphere, offering insights for weather analysis, moisture tracking, and predicting atmospheric instability. Crucial for identifying potential rainfall or thunderstorm development.

Resolution Matters:

Resolution is a pivotal aspect in satellite imagery, determining the quality and quantity of the captured images.

1. **Spatial Resolution:**

Refers to the level of detail captured in an image, determining the smallest distinguishable object. Higher spatial resolution allows for the identification of smaller objects, providing a more detailed representation of the Earth's surface.

1. **Spectral Resolution:**

Relates to a sensor's capacity to discern and capture specific wavelength intervals within the electromagnetic spectrum. High spectral resolution provides detailed information about the Earth's surface composition.

3.**Multispectral and Hyperspectral Sensors:**

Multispectral sensors capture data in specific spectral bands, enhancing our understanding of the environment. Hyperspectral sensors collect detailed information about materials and substances across numerous narrow spectral bands.

Temporal Resolution:

Temporal resolution denotes the frequency at which a satellite captures data for a specific location over time, indicating how often the same area is revisited.

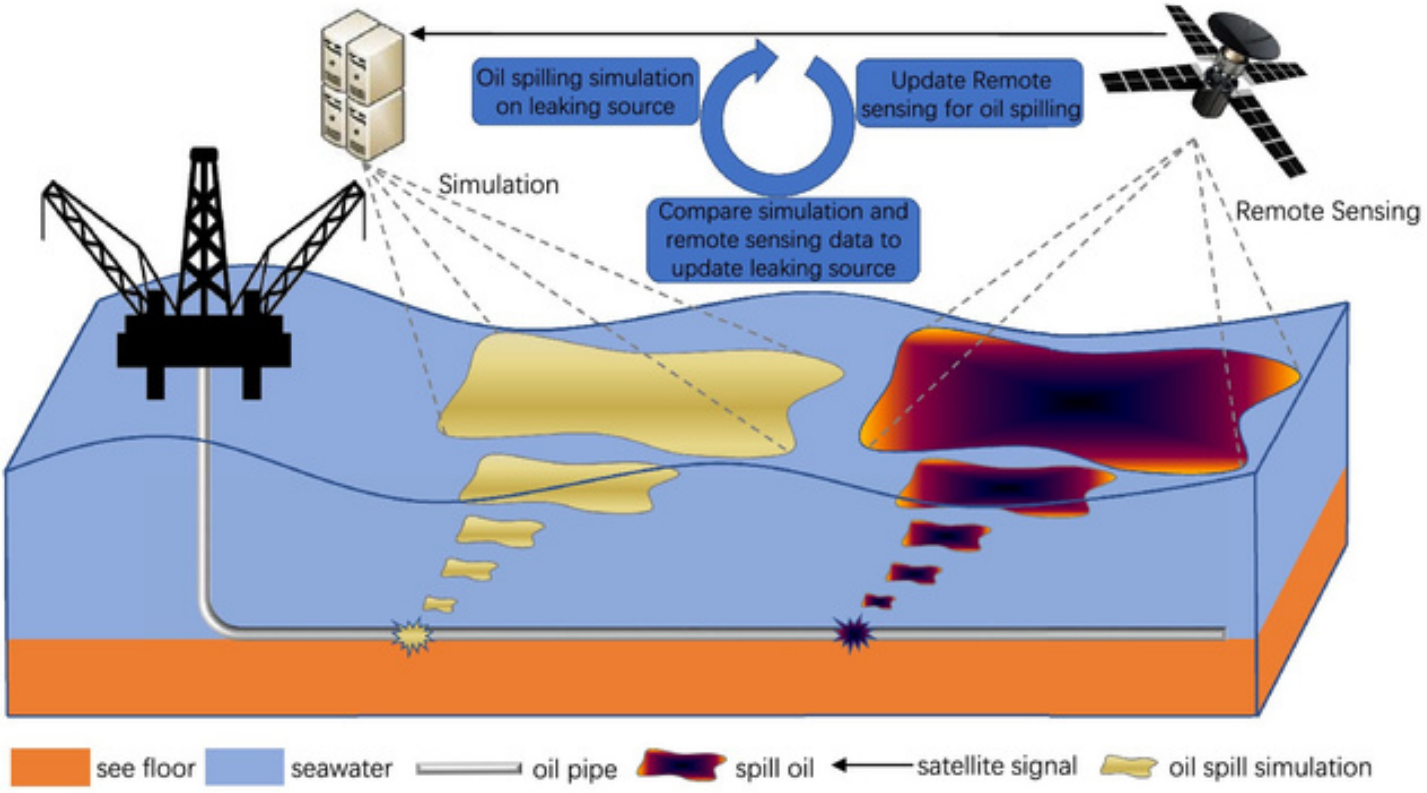
Importance of Temporal Resolution:

Crucial for monitoring dynamic processes such as land cover changes, vegetation growth, urban development, wildfires, or floods.

Radiometric Resolution:

Radiometric resolution refers to a sensor's sensitivity to variations in electromagnetic radiation intensity received from the Earth's surface, indicating the number of grayscale levels that can be imaged.

figure3: Cyber-physical oil spill monitoring and detection for offshore petroleum risk management service



2.3 Satellite Imagery: Landsat

The Landsat program, initiated in the 1960s under Secretary Stewart Udall, has been a pioneering force in the realm of remote sensing. Encompassing a series of Earth observation satellites known as the Landsat Missions, this program, now boasting eight operational satellites, has played a pivotal role in providing valuable data and images of the Earth's surface. These data, originating from the launch of Landsat 1 (ERTS-1) in 1972, were groundbreaking for their ability to observe changes wrought by various forms of human activity. From the resolution that can easily identify larger roads, buildings, and urban development patterns globally to the continuous collection of data, Landsat allows multidecade comparisons, revealing landscape transformations due to activities like urbanization and "slash-and-burn" agricultural practices. Over time, the program has faced challenges, including commercialization efforts in the 1980s, but recognition of the data's value led to its reacquisition by the government, resulting in the more accessible Landsat 7 data. Landsat 5, setting a world record as the longest-operating Earth observation satellite, operated for an impressive 28 years, with Landsat 8 successfully taking the reins following its decommissioning. Despite some technical challenges, the Landsat program's

enduring legacy has revolutionized our understanding of the Earth's surface and continues to be integral in monitoring environmental changes worldwide. The upcoming Landsat Next mission in 2030 promises further advancements, introducing a constellation of three observatories for enhanced temporal revisit, expanded spectral bands, and increased data collection capacity.

2.4 Satellite Imagery: sentinel

The Sentinel program, a cornerstone in contemporary Earth observation, is jointly managed by the European Space Agency (ESA) and the European Commission, offering a comprehensive fleet of satellites equipped with advanced radar and optical imaging technologies. Commencing with the launch of Sentinel-1A in 2014, these satellites have made significant contributions to various fields, including environmental monitoring, climate change analysis, and disaster response. The program's open data policy ensures global access to high-quality imagery for research and operational purposes.

ESA, in conjunction with the Copernicus program, is developing a new family of missions known as Sentinels, specifically designed to meet the operational needs of Copernicus services. These Sentinels consist of satellite constellations carrying advanced technologies, including radar and multi-spectral imaging instruments. The Sentinels play a vital role in environmental and climate monitoring, offering robust datasets for Copernicus by fulfilling revisit and coverage requirements. Specific missions, such as Sentinel-1 for radar imaging, Sentinel-2 for high-resolution land monitoring, and Sentinel-3 for ocean and atmospheric monitoring, cater to distinct services within Copernicus, providing essential data for policymakers and land managers. Additionally, the program's user-friendly data access and extensive coverage make Sentinel a cornerstone in addressing contemporary environmental challenges and enhancing our understanding of the dynamic processes shaping the Earth's surface. Both the Sentinel program and ESA's Sentinel Expansion missions, aligning with EU policy requirements, collectively contribute to a sustained and advanced Earth observation infrastructure.

1. **5-limitations of remote sensing data:**

In the realm of remote sensing, the orchestration of data encounters its set of limitations—a symphony where spatial and temporal resolutions harmonize with atmospheric nuances, spectral paucities, and the cadence of cost and accessibility challenges. Sensory intricacies and the labyrinth of vegetation and topography stand as artistic obstacles, painting the canvas of interpretation with complexity. Processing, akin to crafting a masterpiece, unveils its own challenges. These constraints, though formidable, are integral notes in the melody of data interpretation. Users, as the virtuosos of this grand composition, find it paramount to attune themselves to these nuances, understanding that amidst limitations, the symphony of remote sensing unfolds its richness for diverse applications.

3-Bridging the Gap: Integrating GIS and Geomatics

In the never-ending quest to find new stuff we need; a wide range of tools and techniques is needed to maximize success. This section explores how geographic information systems (GIS) seamlessly integrate with geoscience and remote sensing data to create a more efficient advanced research practices

3.1-Unveiling the Power of GIS:

GIS plays a pivotal role in resource exploration by acting as a central repository and analysis platform for various types of data. It acts as a digital bridge, connecting the dots between:

Remote sensing data: Satellite imagery, radar data, and other remotely sensed information can be integrated into a GIS platform for analysis and visualization.

Geospatial data: Existing datasets, such as geological maps, topographic data, and infrastructure information, can be incorporated into the GIS for comprehensive analysis.

Fieldwork data: Data collected during field surveys, including GPS coordinates, physical samples, and observations, can be integrated for further analysis and visualization.

1. **Understanding the Role of Geomatics:**

Geomatics encompasses a diverse range of techniques and technologies utilized to acquire, manage, analyze, and interpret spatial data related to the Earth. It is vital in resource exploration by providing:

Surveying: Traditional and advanced surveying techniques, like GPS and LiDAR, allow for the precise measurement of the Earth's surface, facilitating the identification of potential resource locations and providing critical data for further analysis.

Seismic Surveys: Subsurface exploration techniques, such as seismic surveys, offer valuable insights into the Earth's geological structure, aiding in identifying potential resource deposits hidden beneath the surface.

Geospatial modeling: Creating three-dimensional models of the Earth's subsurface structure can help visualize potential resource locations and guide further exploration activities.

1. **The Symphony of Integration:**

The true power lies in **integrating** these technologies. By combining GIS, remote sensing data, and geomatics information, we can gain a comprehensive understanding of the target exploration area:

Spatial analysis: GIS allows for overlaying and analyzing various datasets, enabling the identification of spatial relationships and patterns that might be indicative of potential resource locations.

Decision-making support: The combined insights from different data sources provide valuable information for informed decision-making throughout the exploration process.

Enhanced visualization: Integrating data into GIS allows for creating visual representations, such as maps and 3D models, facilitating better communication and collaboration between different stakeholders involved in the exploration project.

Conclusion :

The integration of GIS, geomatics, and remote sensing data plays a pivotal role in modern resource exploration practices. This combined approach allows for:

More efficient and targeted exploration efforts.

Improved accuracy and reliability of exploration results.

Enhanced communication and collaboration within exploration teams.

By understanding the individual strengths of each technology and embracing their synergy, we can unlock the full potential of remote sensing data and pave the way for sustainable and responsible resource exploration practices

1. Harnessing the Power of Intelligence: AI/QML in Exploration

Geo AI is a new research and application field combining spatiotemporal big data analysis and artificial intelligence technology These cutting-edge technologies bring unprecedented capabilities to analyze vast datasets, enabling the identification of intricate patterns and trends for efficient and insightful exploration practices.

1. **Unveiling Hidden Insights:**

Old methods often rely on human expertise for data analysis. AI/ML algorithms, however, possess the ability to:

Process vast quantities of data: They can analyze massive datasets from remote sensing, GIS, and other sources, identifying patterns and trends that might be missed by human analysts.

Identify subtle relationships: AI/ML can identify subtle relationships between different data points, providing valuable insights into potential resource locations and geological characteristics.

Automate repetitive tasks: They can automate time-consuming and repetitive tasks like data cleaning and feature extraction, freeing up valuable resources for further exploration activities.

1. **Embracing a Hybrid Approach: Classical vs. Quantum** ML:

Classic:

Classical bits: Operates on bits, which can be either 0 or 1. Think of a light switch, on or off.

Sequential processing: Analyzes data one piece at a time, similar to solving a math problem Limited by complexity: Struggles with highly complex problems with many variables.

Quantum:

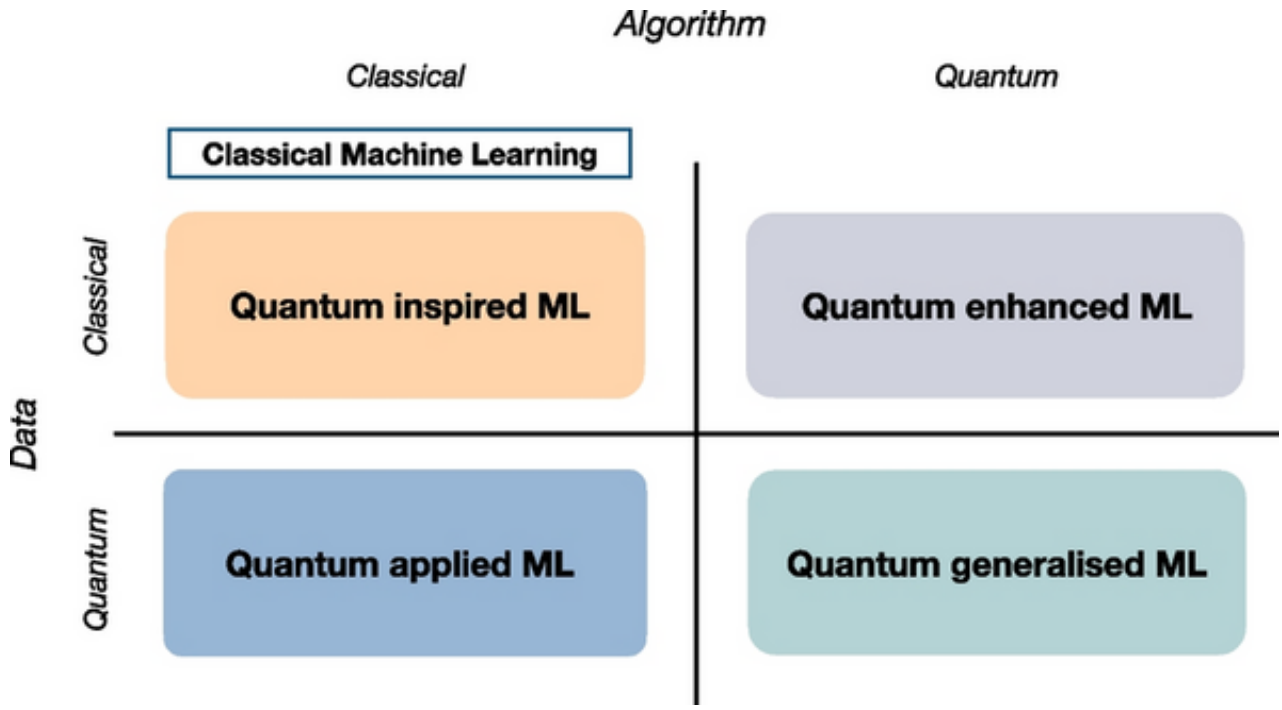
Quantum bits (qubits): Utilizes qubits, which can hold both 0 and 1 simultaneously, a concept known as superposition. Imagine a light switch being both on and off at the same time (not possible in our current understanding of physics, but an important concept in quantum mechanics).

Parallel processing: Explores multiple possibilities simultaneously, like checking all the answers in a math test at once.

Potential for tackling complex problems: Holds promise for solving problems intractable for classical ML due to their inherent complexity.

While QML holds immense potential, full-fledged quantum computers are still under development. Therefore, this project adopts a **hybrid approach** We combine the strengths of traditional ML, which excels at handling large datasets, with the potential benefits of QML for specific tasks like the complexity of the data and the GPU usage

figure4: quantum machine learning development paradigms compared against data and algorithms type



4.4 Synergistic Integration with Geomatics:

Integrating AI/ML and geomatics techniques unlocks a new level of exploration:

Data analysis and interpretation: AI/ML algorithms can analyze geomatics data (e.g., seismic surveys) to identify potential resources or understand subsurface structures.

Enhanced decision-making: Combining AI/ML insights with geospatial data and geomatics interpretations leads to informed exploration decisions.

1. The Synergy in Action: A Collaborative Approach:

5.1 Orchestrating the Symphony: A Combined Workflow

Our project's exploration workflow thrives on the **synergy** between various technologies:

1. **Remote sensing data acquisition:** We utilize satellite imagery (Landsat, Sentinel) to capture high resolution images of the Kansas City.

1. 1. **GIS analysis:** We integrate the acquired data into a GIS platform, combining it with existing geospatial datasets and field survey data.
2. 1. **Geomatics exploration:** We leverage geomatic techniques like LiDAR to gather high-precision elevation data and potentially seismic surveys to probe deeper into the subsurface structure.
3. 1. **AI/QML integration:** We employ AI/QML algorithms to analyze the combined data, searching for hidden patterns.

1. **5.2 A Collaborative Advantage: A Hypothetical Example**

Imagine exploring a vast desert region for potential mineral deposits. Traditionally, geologists might rely on:

Visual inspection of satellite imagery: This approach is time-consuming and prone to overlooking subtle anomalies.

Limited field surveys: Ground surveys are expensive and can only cover a small fraction of the area.

Our combined approach offers distinct advantages:

- We can analyze vast imagery datasets, identifying potential areas of interest based on spectral signatures or subtle geological features.
- By overlaying geological maps and existing exploration data in GIS, we can prioritize these areas for further investigation
- Utilizing LiDAR data, we can create detailed terrain maps, guiding field surveys to the most promising locations identified through AI/QML analysis
- Seismic surveys, combined with AI/QML analysis, can provide valuable insights into potential subsurface structures indicative of mineral deposits
- By focusing on high-potential areas, we can conduct fewer but more targeted field surveys.

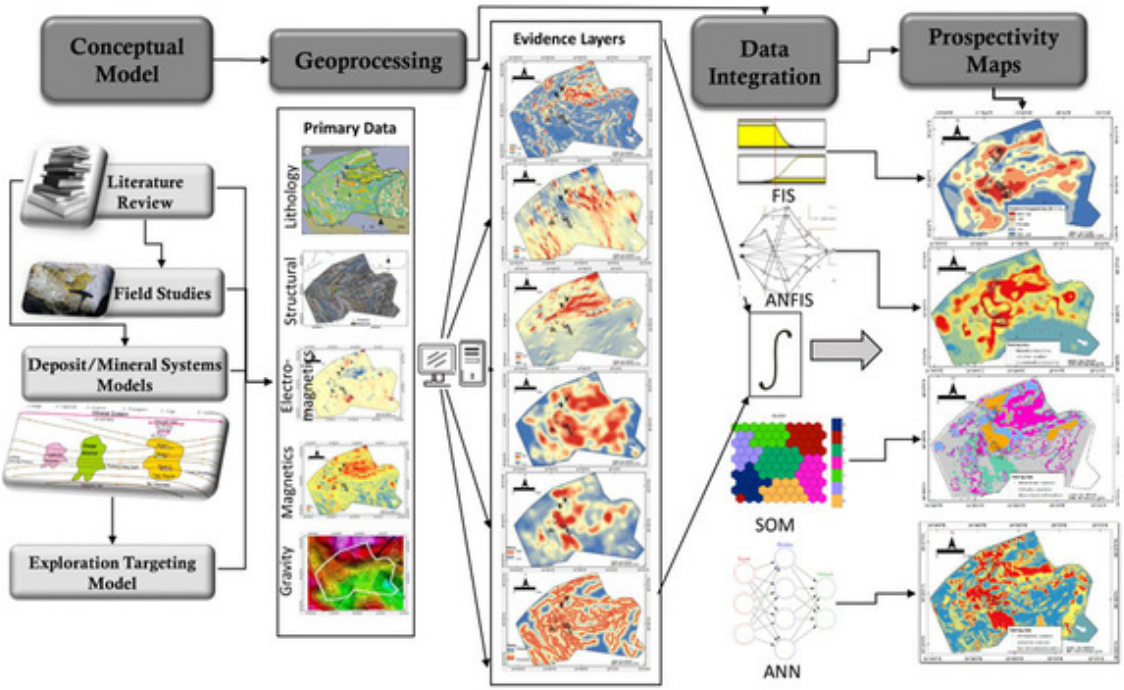


figure5: Illustration of a Collaborative Exploration Workflow for Desert Mineral Deposit

5.3 Conclusion:

The fusion of remote sensing, GIS, geomatics, and AI/QML, we unlock a powerful and efficient approach for resource exploration. This collaborative approach makes a roadmap for a future of sustainable and responsible resource discovery.

Chapter 4: Study conception from pixels to predictions

Introduction:

In this chapter, we are going to dig deeper into the roadmap of our AI/QML-driven resource exploration project. Our primary objective is to outline the workflow, detailing the specific software tools we will utilize and the different types of data we will employ. development platform and model, the data acquisition process for various sources, and the image processing steps to prepare the data for analysis. By outlining this comprehensive workflow, we aim to establish a solid foundation for building and applying our model for successful search.

1- Building the Foundation: Software Selection:

The building blocks of any study are the proper selection of the right software, and as GIS scientists, we know for fact that ArcGIS and other GIS software are crucial for this. With our finest programs, we begin with ArcGIS.

1a- ArcMap vs ArcGIS Pro :

While both ArcMap and ArcGIS Pro are products from Esri and fall under the ArcGIS suite, they cater to different needs and workflows, but in this project, we used both software's to maximize all the benefits from them.

ArcMap:

Strengths:

Maturity and Stability: ArcMap has been around for a longer time, offering a mature and stable platform with a vast library of extensions and functionalities.

Customization: ArcMap offers a high degree of customization through add-ins and extensions, allowing users to tailor the software to their specific needs.

Weaknesses:

Limited Support: Esri has shifted its focus to ArcGIS Pro, resulting in diminished support for ArcMap.

Future updates and bug fixes might be less frequent.

32-bit Architecture: ArcMap operates on a 32-bit architecture, limiting its ability to handle very large datasets. Or 64-bit float architecture

Limited Python 3 Support: ArcMap's Python scripting capabilities primarily rely on Python 2.7, which is reaching end-of-life.

ArcGIS Pro:

Strengths:

Modern Architecture: ArcGIS Pro leverages a 64-bit architecture, allowing it to handle large and complex datasets more efficiently.

Advanced Functionality: ArcGIS Pro offers a wider range of built-in functionalities compared to ArcMap, including advanced spatial analysis tools and improved 3D visualization capabilities. **Active Development:** Esri actively develops and updates ArcGIS Pro, ensuring access to the latest features and bug fixes.

Python 3 Support: ArcGIS Pro fully supports Python 3, opening doors for leveraging a wider range of Python libraries and functionalities for geospatial analysis and AI/QML integration.

Weaknesses:

Learning Curve: Due to its newer interface and functionalities, ArcGIS Pro might have a steeper learning curve for users accustomed to ArcMap.

Limited Customization: While customization options exist, they are not as extensive as those offered by ArcMap.

While both ArcMap and ArcGIS Pro are valuable tools ArcGIS Pro aligns better with the contemporary approach to geospatial analysis, particularly its integration with Python 3, which is more than necessary for AI/QML development

The 64-bit architecture of ArcGIS Pro facilitates working with more complex dataset like ours, which is likely encountered in resource exploration involving satellite imagery and other geospatial information.

QGIS (Quantum GIS):

Open-source: Free and readily available, making it a cost-effective option for our case as a student or organizations with budget constraints.

Wide Functionality: Offers a comprehensive set of functionalities for GIS data visualization, spatial analysis, and basic image processing.

Python Integration: Similar to ArcGIS Pro, QGIS supports Python scripting, allowing for custom workflows and integration with AI/QML libraries (though potentially requiring more technical expertise compared to ArcGIS Pro).

Strengths:

Strong user community with extensive online resources and tutorials.

Offers plugins for specialized tasks, potentially including some related to AI/QML (although these might be less developed compared to commercial software).

Weaknesses:

Limited native support for advanced image processing tasks often encountered in resource exploration (compared to specialized software like ENVI).

require more technical expertise to set up and customize workflows compared to user-friendly commercial options.

ENVI (Environment for Visualizing Images):

Commercial Software: Paid software with a licensing fee, offering a range of functionalities tailored for remote sensing image processing and analysis.

Advanced Image Processing: ENVI excels in advanced image processing tasks like atmospheric correction, spectral band manipulation, and feature extraction, which is exactly what we need in this particular project for preparing remote sensing data for resource exploration applications.

Specialized Workflows: Offers pre-built workflows and tools designed specifically for resource exploration tasks, potentially including mineral mapping or hydrocarbon exploration.

Weaknesses:

Cost: The commercial license can be expensive, especially for individual users or smaller research projects.

Slower Learning Curve: The extensive functionalities might require more time and effort to master compared to user-friendly GIS platforms like ArcGIS or QGIS.

MAGMAP (MAGnetic MAPping):

Commercial Software: MAGMAP is a paid software solution with a licensing fee, catering to professionals and organizations involved in geophysical exploration and research. Its pricing model typically includes commercial licenses, which can vary depending on the scale of usage and specific requirements.

Advanced Magnetic Data Processing: MAGMAP stands out for its advanced capabilities in magnetic data processing and interpretation. It offers a suite of tools and algorithms tailored to handle magnetic data effectively. This includes functionalities such as data visualization, filtering, modeling, and interpretation, essential for analyzing our magnetic anomalies and identifying potential subsurface structures. We used it to manipulate our magmatic map and the LAS file

Specialized Workflows: MAGMAP provides specialized workflows and tools explicitly designed for resource exploration tasks, particularly in the field of geophysics. These workflows are optimized for magnetic data interpretation and analysis, covering various exploration activities such as mineral mapping and hydrocarbon exploration.

By offering pre-built workflows and specialized tools, MAGMAP streamlines the exploration process, enabling efficient data analysis and decision-making.

Weaknesses:

Cost: One of the primary drawbacks of MAGMAP is its cost. The commercial license is very expensive, Organizations or research teams considering MAGMAP may need to allocate sufficient resources to cover the licensing fees.

Visual Studio Code (VS Code):

Open-source and Free: VS Code is free to use and open-source,

Extensive Language Support: Offers support for numerous programming languages, including Python, R, which is widely used in GIS scripting and automation.

Customizable: VS Code is highly customizable with a vast library of extensions, allowing users to tailor their development environment to their specific needs.

Integration with Git and copilots: Built-in Git support enables version control and collaboration on geospatial projects and chat copilots helps with coding.

- **Well-suited for Scripting and Development:** VS Code is particularly well-suited for scripting, development, and integrating with other tools and libraries commonly used in GIS workflows, such as GDAL and geospatial Python libraries.

Lacks Built-in GIS Functionality: While VS Code provides a versatile development environment, it lacks built-in GIS functionality for spatial analysis, visualization, and data processing compared to dedicated GIS software like ArcGIS and QGIS.

Requires Additional Extensions: installing extra extensions or plugins to add GIS-related functionalities is needed, which may not always provide the same level of integration and ease of use as dedicated GIS software.

Conclusion:

a variety of software platforms cater to different needs and project requirements. Having experience with ArcGIS, **QGIS**, and **ENVI** allows for a flexible approach.

ArcGIS Pro remains the primary platform for this project due to its strengths in

- Python 3 support simplifies
- Data Integration

Qgis is the secondary platform for exceling at:

- **Open source**
- **Python scripting**

Envi was the best when it comes to

- feature extraction such as performing liniment analyze
- Spectral band manipulation

This section acknowledges the value of all six software's, from data acquisition and processing to model development and analysis. It also showcases my expertise in working with all six platforms and explains how each can contribute to the overall workflow. Now that we have covered all the software we need, let's move a step forward into data analysis and types.

2- Data Acquisition: Gathering the Raw Materials:

This section of Chapter 4 will cover the strategy for acquiring the different data types essential for an AI/QML driven resource exploration project. Here's a breakdown of the key points:

2-1. Remote Sensing Data :

USGS Earth Explorer https://earthexplorer.usgs.gov/

Given its extensive data archive and free access, the Landsat program offered by USGS presents a valuable resource for our remote sensing data acquisition. We will explore the available Landsat imagery for our target area (Kensas), focusing on selecting spectral bands that effectively detect hydrocarbons. This selection will consider the trade-off between spatial resolution (30 meters) and data volume. The downloaded Landsat data, in a format compatible with our GIS platform, will form a key component of our AI/QML model development.

Landsat 9 :

Launched on September 27, 2021, Landsat 9 is the newest satellite in the Landsat program, which has been providing valuable Earth observation data since the early 1970s. Landsat satellites take high-resolution pictures of the Earth's

surface, which are utilized for a variety of applications such as agriculture, urban planning, forestry, environmental monitoring, and disaster relief.

The explanation of Landsat 9 and the factors that make it superior to Landsat 8 is provided below:

Better Sensor Technology: The Thermal Infrared Sensor 2 (TIRS-2) and Operational Land Imager 2 (OLI-2) in Landsat 9 are better sensors than those on Landsat 8. These sensors' improved spectral, radiometric, and spatial capabilities result in crisper, more detailed images of the Earth's surface.

Continuity of Data: Landsat 9 ensures the continuity of the Landsat program's data record, which is crucial for monitoring long-term changes in Earth's environment. By providing consistent, high-quality imagery, Landsat 9 contributes to ongoing research and applications in areas such as land use planning, natural resource management, and climate change monitoring.

Extended Lifespan: Landsat 9 is designed to operate for at least five years, with the potential for an extended mission lifespan. This ensures a reliable and continuous stream of data for the scientific community, allowing for the monitoring of both short-term events and long-term trends.

Overall, Landsat 9 represents a significant advancement in Earth observation technology, building upon the success of previous missions like Landsat 8. With its enhanced sensors, improved **spatial** and **temporal** resolution, and **continuous data,** Landsat 9 **will provide** researchers and **decision makers with** valuable **insight** into the dynamic processes **that shape the Earth**.

Nine spectral bands:

Band 1 Visible Coastal Aerosol (0.43 – 0.45 µm) 30-m

Band 2 Visible Blue (0.450 – 0.51 µm) 30-m

Band 3 Visible Green (0.53 – 0.59 µm) 30-m

Band 4 Red (0.64 – 0.67 µm) 30-m

Band 5 Near-Infrared (0.85 – 0.88 µm) 30-m

Band 6 SWIR 1(1.57 – 1.65 µm) 30-m

Band 7 SWIR 2 (2.11 – 2.29 µm) 30-m

Band 8 Panchromatic (PAN) (0.50 – 0.68 µm) 15-m

Band 9 Cirrus (1.36 – 1.38 µm) 30-m

Thermal Infrared Sensor 2 (TIRS-2)

Landsat 9's Thermal Infrared Sensor 2 (TIRS-2) measures thermal radiance emitted from the land surface in two thermal infrared bands using the same technology that was used for TIRS on Landsat 8, however TIRS-2 is an improved version of Landsat 8's TIRS, both with regards to instrument risk class and design to minimize stray light. TIRS-2 provides two spectral bands with a maximum ground sampling distance, both in-track and cross track, of 100 m (328 ft) for both bands. TIRS-2 provides an internal blackbody calibration source as well as space view capabilities. TIRS-2 is designed by NASA Goddard Space Flight Center in Greenbelt, Maryland.

Two spectral bands:

Band 10 TIRS 1 (10.6 – 11.19 µm) 100-m

Band 11 TIRS 2 (11.5 – 12.51 µm) 100-m

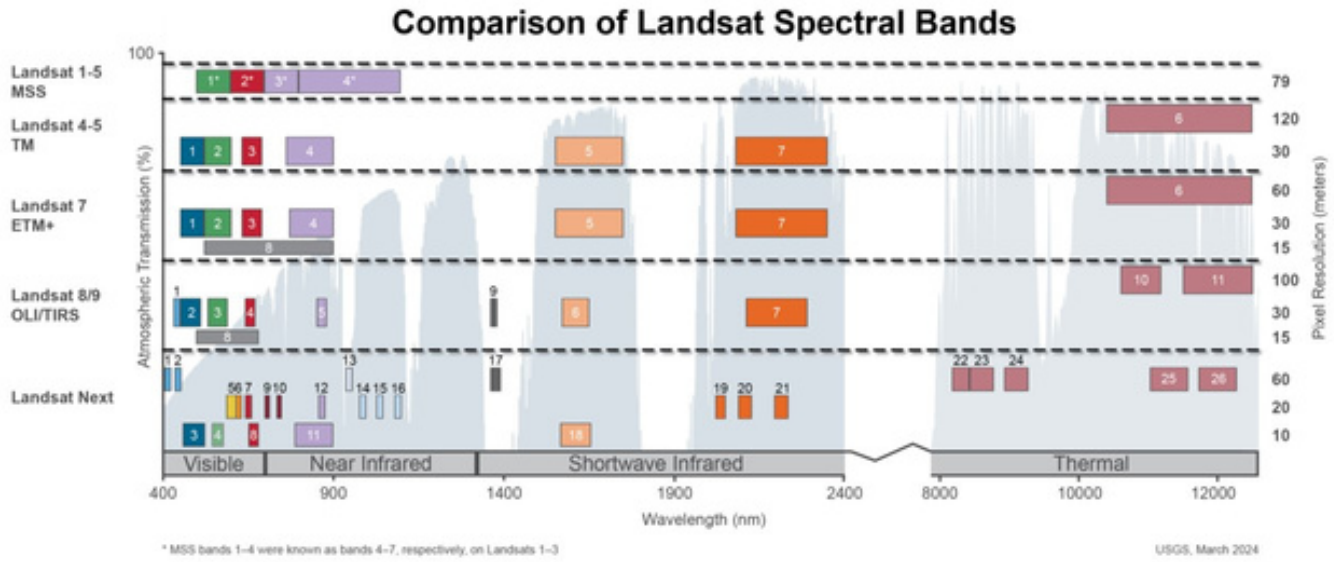


Figure6: Landsat Spectral Bands

2-2 Region of interest (Kansas City):

Nestled at the confluence of the Kansas and Missouri Rivers, Kansas City emerges as a vibrant metropolis steeped in rich history, cultural diversity, and economic significance. Situated at approximately 39.0997° N latitude and 94.5786° W longitude, the city spans across the borders of both Kansas and Missouri, serving as a beacon of Midwest charm and dynamism. But beyond its bustling urban landscape lies a lesser-known aspect of Kansas City's past and present – its role in the exploration and extraction of oil and gas resources.

Kansas City's location in the center of the United States indicates its strategic location in the region. The city's location near the geographic center of the American continent indicates its vital role as a crossroads of transportation, commerce, and cultural exchange. These coordinates not only indicate your physical location, but are also gateways to the great center of the United States.

The latitude of Kansas City, Missouri, USA is 39.099724 and the longitude is -94.578331. Kansas City, Missouri, United States is located in the United States in the urban areas zone with GPS coordinates 39° 5' 59.0064" N and 94° 34' 41.9916" W

Kansas City's oil and gas history dates back to the late 1800s, when miners began tapping the vast underground reserves beneath the fertile plains of the Midwest. Originally an oil finds in neighboring states such as Texas and Oklahoma, Kansas City soon became a center for exploration and drilling.

One of the region's earliest and most important oil discoveries was made near the city in 1892. south of Paola, Kansas City. The resulting oil boom not only transformed the local economy, but also spurred rapid urbanization and infrastructure development in and around Kansas City.

Kansas City's energy exploration landscape has evolved over the decades, and advances in technology have led to more efficient extraction methods. and increased production rates. Today, the city and surrounding areas are dotted with oil wells, platforms and refineries, a constant reminder of the region's rich oil heritage.

Due to its deep ties to the energy industry, Kansas City has also embraced renewable energy sources. in recent years, reflecting the growing global trend towards sustainable development and environmental protection. Wind farms, solar farms and biofuel plants have become increasingly common sights in the region, signaling a shift toward cleaner, more sustainable energy options.

In addition to its role in energy exploration, Kansas City continues to attract visitors with its own. vibrant cultural life.

a world-famous barbecue and thriving arts community. From the iconic fountains of Country Club Plaza to the historic jazz clubs at 18th and Vine, the city offers residents and tourists a variety of attractions and experiences.

As we look to the future, Kansas City remains a dynamic and evolving landscape shaped by its past with an eye to tomorrow. Whether you explore its bustling city streets or venture into the peaceful countryside, one thing is for sure, this Midwestern gem continues to inspire and fascinate all who encounter its boundless energy

2-3 -2 Well Logs & Geologic Information: Subsurface Insights:

This subsection highlights the importance of well logs and geologic information for the project. with a breakdown of the key points.

Data Description:

Well logs are digital information compiled in the course of the drilling procedure of oil and fueloline wells, supplying important insights into subsurface situations. They embody numerous parameters such as:

Depth Measurements:

Records the intensity at which every information factor is measured alongside the wellbore, facilitating unique evaluation and correlation of geological features.

Lithology:

Describes the varieties of rocks encountered in the course of drilling, helping in information the geological formations and predicting reservoir characteristics.

Porosity:

Indicates the share of void areas inside the rock formation, important for assessing the reservoir‘s cap potential to save hydrocarbons.

Permeability:

Reflects the rock's cap potential to permit fluid waft thru its pore areas, influencing the convenience with which oil, fueloline, or water can circulate inside the reservoir.

Fluid Content:

Identifies the presence and distribution of various fluids (e.g., water, oil, fueloline) inside the formation, critical for estimating hydrocarbon reserves and making plans manufacturing strategies.

These information factors are pivotal for reservoir characterization, formation evaluation, and decision-making approaches in oil and fuel line exploration and manufacturing operations.

Well log’s function vital equipment for geoscientists and engineers to recognize subsurface situations and optimize drilling and manufacturing strategies.

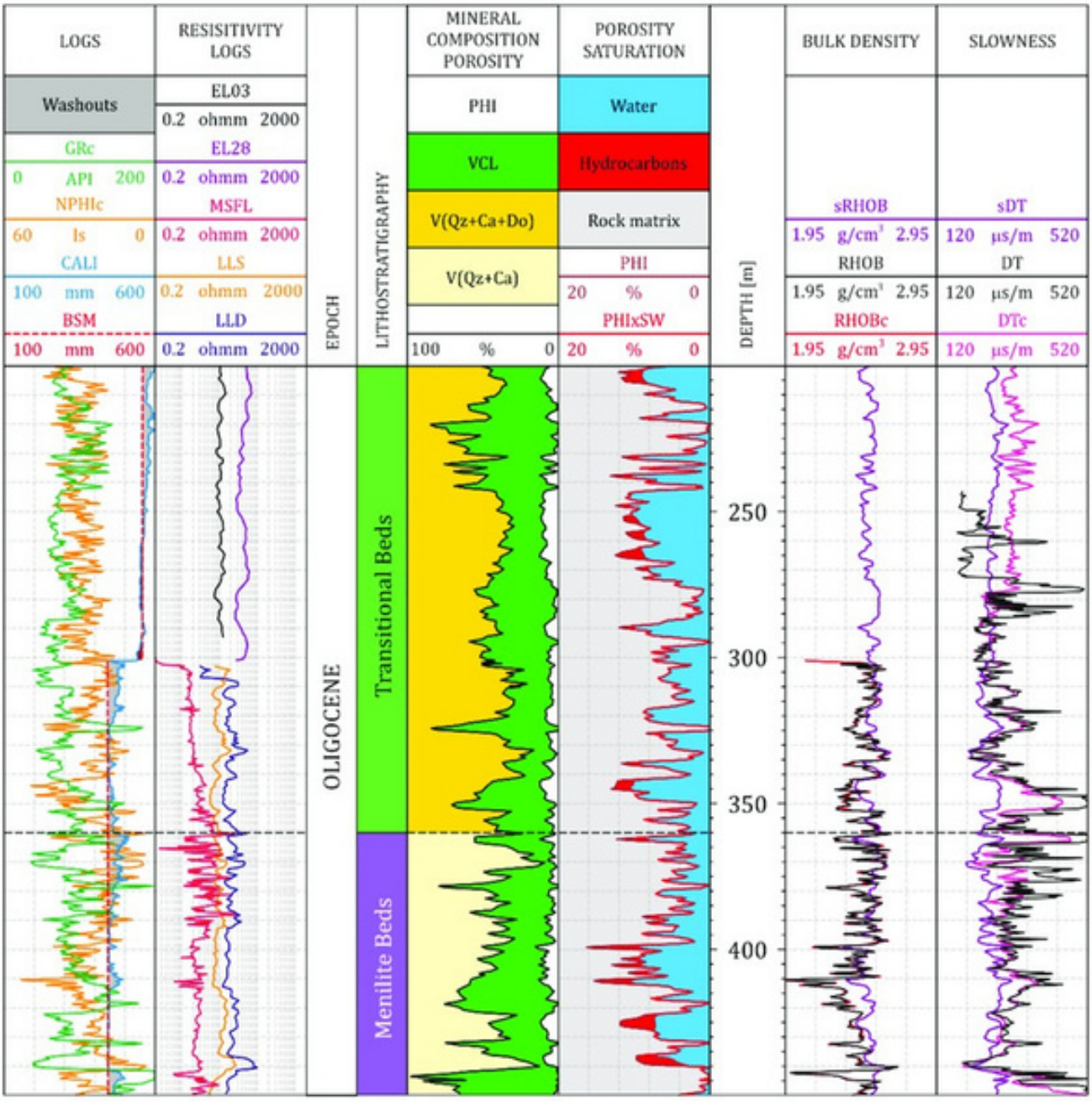


Figure7: diagram of Well logging data in the uppermost depth section of the D-1 borehole

Geological information plays a pivotal role in hydrocarbon exploration, particularly in regions like Kansas City. where understanding the subsurface environment is essential.

Here's a manifestation of well logs and geological data in this context:

- Identification of Potential Reservoirs: This kind of data provides accurate information about the geological formations present in our region of interest. Analyzing the data helps us identify potential hydrocarbon reservoirs based on the presence of the right type of rock formations. These formations may include the right combination of characteristics, such as porosity, permeability, and even the location.
- Delineation of Trap Structures: Some structures, like anticlines, faults, and stratigraphic traps, can act as traps for hydrocarbons. With the right treatment of geological maps and seismic data, geologists can identify these structural

features and assess their potential to trap hydrocarbons. Well logs then provide detailed information about the geometry and characteristics of these traps, which is important for their identification.

- Depth and Thickness: Well logs provide data on the depth and thickness of different geological layers, including potential mineral-bearing formations. helping in the planning and execution of drilling operations.
- Validation Using Historical Well Data: Geological interpretations and predictions can greatly benefit from the use of historical well data from earlier drilling operations in the region. Geologists can analyze and correlate geological formations in various areas using well logs from existing wells, which enhances the precision of subsurface models and predictions.

In conclusion, well logs and geological information are indispensable tools for understanding the subsurface. geology and identifying potential hydrocarbon reservoirs in regions like Kansas City.

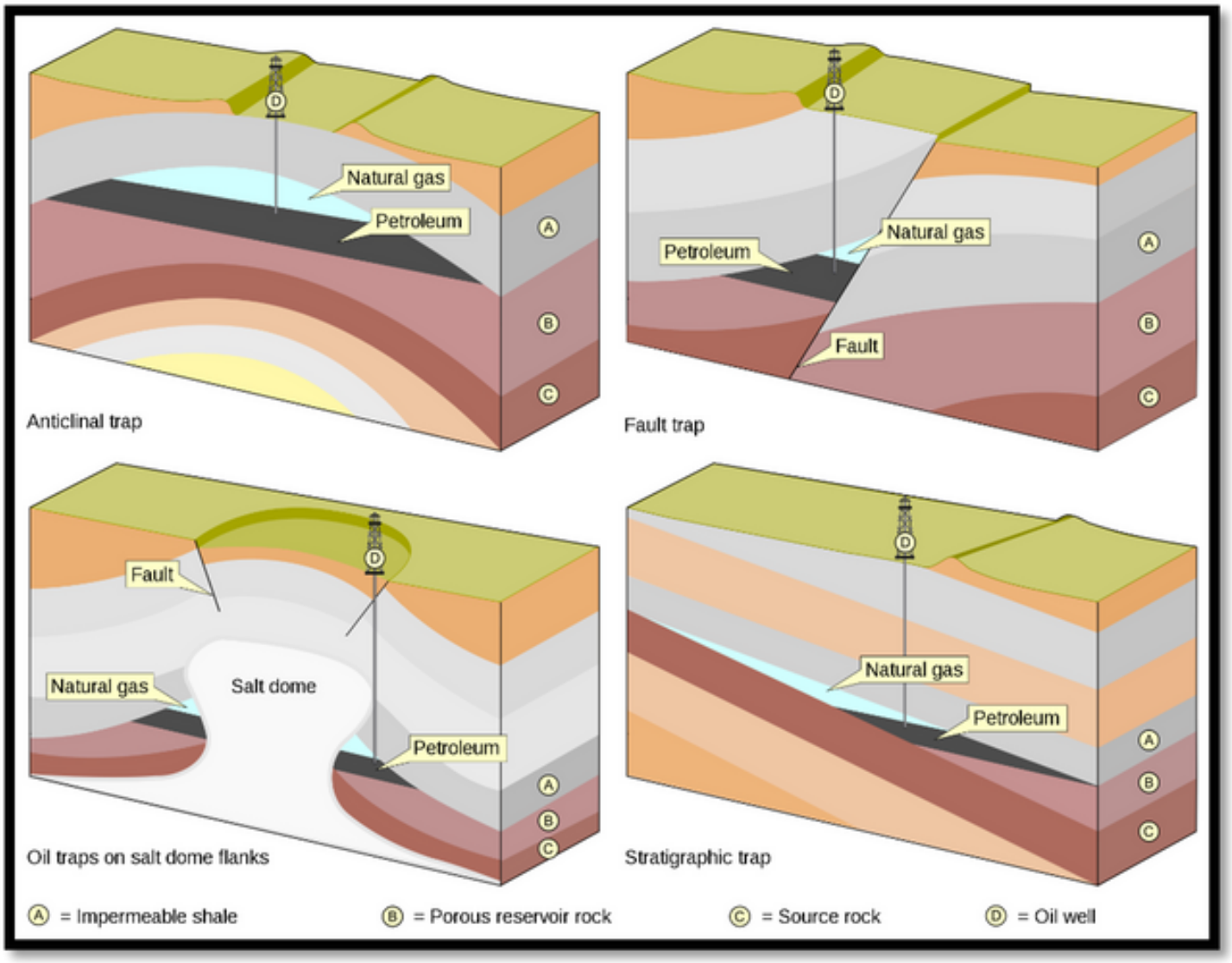


figure 8: Diagram showing the structure of several different types of oil and gas traps

2.3.3 LAS File & Seismic Data: Exploring Deeper:

In addition to well logs and geologic information, our exploration strategy incorporates LiDAR (Light Detection and Ranging) data stored in LAS format and seismic data. we aim to develop a comprehensive understanding of the subsurface structure within our target area (Kansas City, USA). This comprehensive approach will provide valuable input for our QML model, allowing it to analyze a broader range of features associated with hydrocarbon exploration.

Originally, lidar data was only delivered in ASCII format. With the massive size of lidar data collections, a binary format called LAS was soon adopted to manage and standardize the way in which lidar data was organized and disseminated. Now

lidar data is commonly represented in LAS. LAS is a more acceptable file format, because LAS files contain more information and, being binary, can be read by the importer more efficiently.

LiDAR (LAS File):

LiDAR technology employs light pulses emitted from a laser sensor to accurately measure distances to the Earth's surface. These pulses generate high-resolution 3D models known as LAS files. This data is valuable for several reasons:

Identifying surface features potentially indicative of subsurface structures: LiDAR data can reveal subtle variations in terrain elevation, allowing geoscientists to identify surface expressions of underlying geological features such as faults, folds, and other structural complexities.

Generating detailed topographic maps: LiDAR captures precise elevation data, enabling the creation of highly detailed topographic maps that are essential for understanding surface morphology and landscape evolution.

Seismic Data:

Seismic data acquisition involves generating controlled sound waves (usually using specialized equipment such as seismic vibrators or explosives) and recording their reflections from subsurface rock layers. Geophysicists analyze these reflections to construct detailed images of the subsurface. Seismic data can reveal:

Faults and folds in rock formations: By analyzing the patterns of seismic reflections, geoscientists can identify faults—fractures in the Earth's crust where movement has occurred—and folds—bends or wrinkles in rock layers caused by tectonic forces.

Potential hydrocarbon reservoirs: Seismic surveys are widely used in oil and gas exploration to identify underground structures that may contain hydrocarbon reservoirs. Specific seismic reflections can indicate the presence of porous rock formations that may trap oil or gas.

Value	Meaning
0	Created, Never classified
1	Unassigned
2	Ground
3	Low Vegetation
4	Medium Vegetation
5	high Vegetation
6	Building
7	Low Point
8	Model Key–Point
9	Water
10	Rail
11	Road Surface
12	Reserved
13	Wire – Guard (Shield)
14	Wire – Conductor(Phase)
15	Transmission Tower

16	Wire-Structure Connector
17	Bridge Deck
18	High Noise
19	Reserved
20	Ignored Ground
21	Snow
22	Temporal Exclusion
23-63	reserved
64-255	User Definable

Table 1: Classification codes for LAS formats 1.1 through 1.4

2.3.4 magnetic data integration:

Magnetic data, acquired through geophysical techniques like aeromagnetic surveys, can also be integrated into our analysis. Aeromagnetic surveys involve flying low-altitude aircraft equipped with magnetometers to measure variations in the Earth's magnetic field caused by subsurface rocks. Geologic structures like faults and ore bodies containing iron-rich minerals can cause disruptions in the magnetic field, creating measurable anomalies. By processing and analyzing this magnetic data, geophysicists can create digital aeromagnetic maps that reveal these anomalies.

3- Data preprocessing : preparing the canvas

Before jumping into the code, we need to meticulously prepare our gathered data and perform a tedious amount of tasks to make it clear and ready. In this section, we will take a journey through the process.

3.1 collection and preprocessing :

The collection of the various data took a long time, and for this kind of work, Google Search Engine was not enough. We needed to take the search to the next level by launching a new Web 2.0 search engine called Explorer.

https://explorer.globe.engineer

Given the easy access we have to a large amount of data, we were able to access the University of Kansas, which was the biggest source of our data, from geologic maps to well logs. Using a technique called Google Dorking, we were able to

gather most of the information we needed. This technique is to use specific key words to narrow down the research to make it more specific about our interest.

3.1.1 landsat9 :

Landsat 9 is the most suitable option, as we previously discussed we downloaded the bands and the amount of data was enormous so we had to buy time with our power-shell to create multiple folders with one command **for ($i=1;$i -le 10;$i++){New-Item -ItemType Directory -Path ".\Folder$i"}**

The first step we took was to organize the images

now the Preprocessing Landsat 9 imagery with ArcGIS typically involves several steps to correct atmospheric, geometric, and radiometric distortions. Here's a brief description of each step:

Enhancing the resolution:

To enhance the spatial resolution of Landsat 9 imagery from its native 30-meter resolution to 15 meters, several techniques can be employed. One common method is through image fusion or pan-sharpening.

Pan-sharpening involves combining the higher-resolution panchromatic (black and white) band with the lower resolution multispectral bands to create a single high-resolution color image. This process effectively increases the spatial detail of the multispectral bands while preserving their spectral characteristics.

In ArcGIS, pan sharpening tools are available to perform this enhancement. These tools use algorithms such as Bovey, Gram-Schmidt, or Principal Component Analysis (PCA) to fuse the panchromatic and multispectral bands. The result is a higher-resolution image that retains both the spatial detail of the panchromatic band and the spectral information of the multispectral bands.

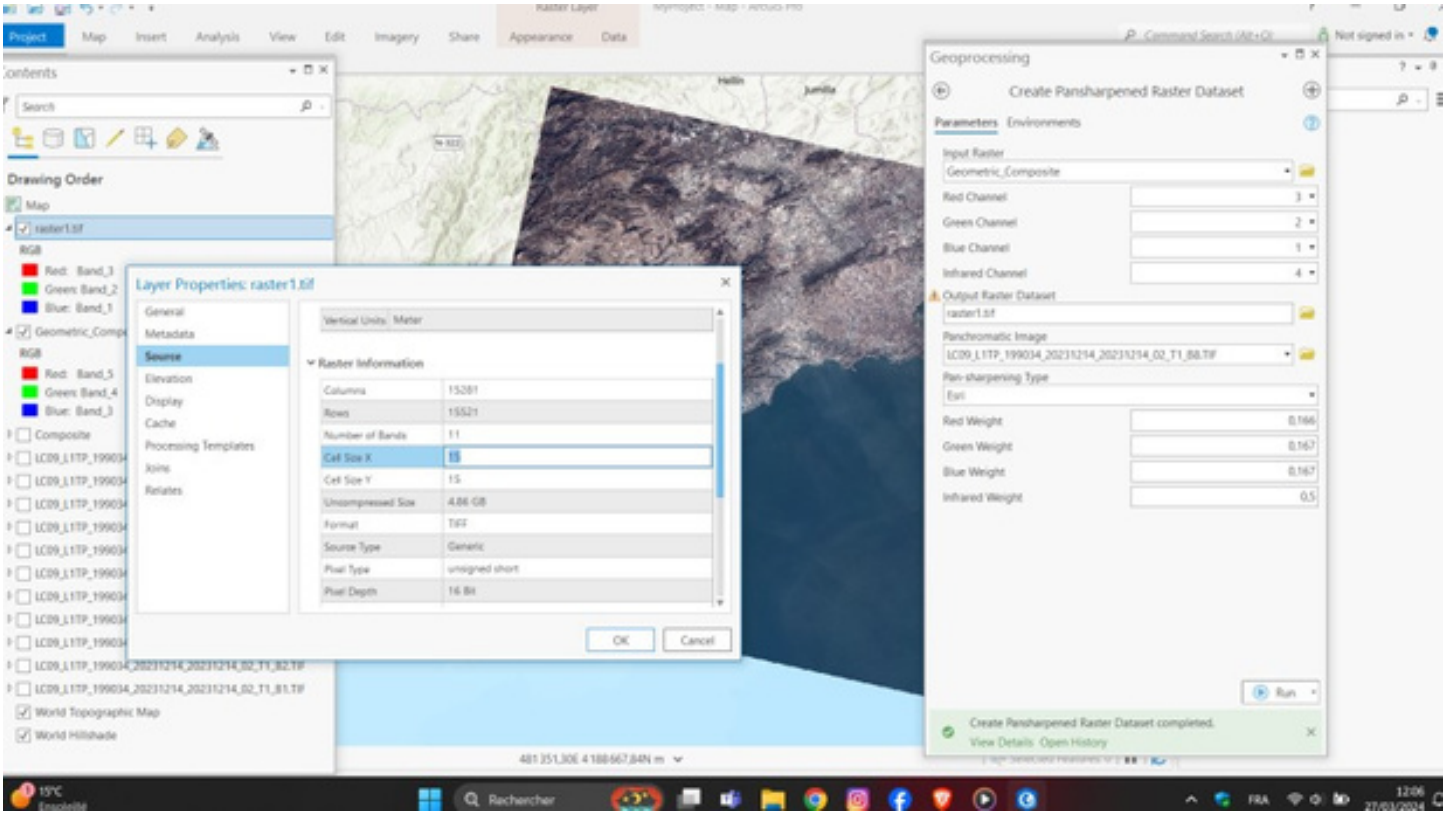


figure 9: enhancing resolution in ArcGIS pro

Geometric Correction: Landsat imagery often suffers from geometric distortions due to factors like terrain variations and satellite sensor characteristics. Geometric correction involves registering the image to a known coordinate system, correcting for distortions caused by the Earth's surface. This is typically done using ground control points (GCPs) and resampling techniques to align the image with a map projection system.

Radiometric Correction: Radiometric correction aims to adjust pixel values to remove distortions caused by atmospheric conditions, sensor characteristics, and solar illumination angles. This correction involves converting digital number (DN) values to radiance or reflectance values, which are more consistent across different images and can be directly compared for analysis.

Atmospheric Correction: Atmospheric correction is influential for removing the effects of atmospheric scattering and absorption, which can distort the spectral characteristics of the image. This correction typically involves applying models to estimate and remove atmospheric effects, such as Rayleigh scattering, aerosol scattering, and water vapor absorption This procedure also involves the conversion of raw spectral band measurements into reflectance values, facilitating more meaningful quantitative analyses and interpretations of the observed environmental phenomena.

REFLECTANCE_MULT_BAND_1 = 2.0000E-05

REFLECTANCE_ADD_BAND_1 = -0.100000

then we multiply the reflectance mult band by the band2 and divide it by the sin of the sun elevation in radiance unit.

and for this we used ArcMap for more efficiency

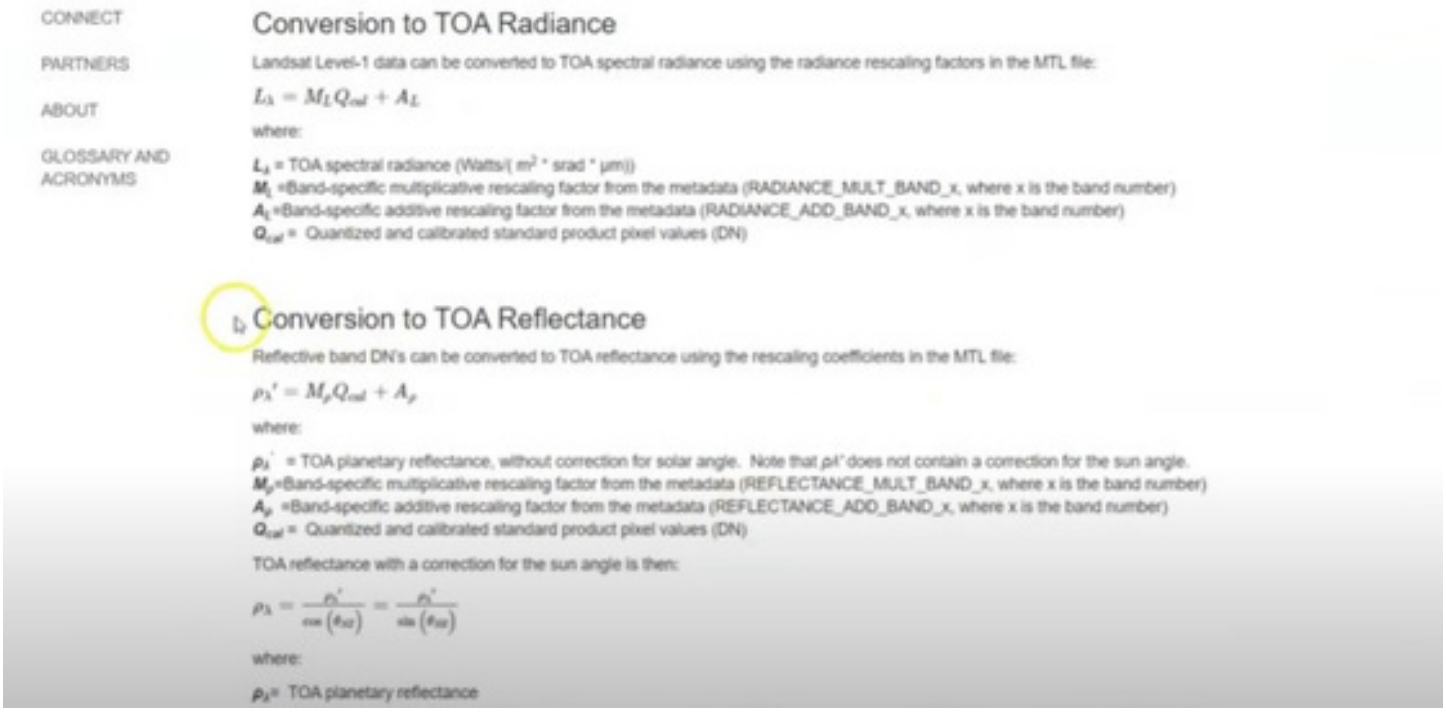


figure 10: USGS figure of toa values and conversion

Mosaicking: If Landsat imagery consists of multiple scenes covering the same area, mosaicking is performed to create a seamless composite image. This involves blending overlapping regions of adjacent scenes to create a single continuous image for analysis.

Subset and Masking: Sometimes, you may only be interested in a specific region within the Landsat scene. Subset and masking techniques are used to extract the desired region of interest while excluding irrelevant areas.

Band compositing:

Band compositing involves combining individual bands from satellite imagery to create a multi-band composite image. This process enhances the interpretability of satellite data by emphasizing certain spectral bands or features of interest. Mosaicking is the initial step, where multiple scenes covering the same area are seamlessly blended together to form a continuous image. Subset and masking techniques are then applied to isolate specific regions of interest while excluding irrelevant areas, allowing for focused analysis. Following this, enhancement and visualization methods such as contrast stretching, color compositing, and histogram equalization are employed to improve the clarity and highlight key features within the composite image. These preprocessing steps ensure that the composite image is accurate, calibrated, and ready for further analysis and interpretation within GIS platforms like ArcGIS.

Enhancement and Visualization: After preprocessing, Landsat imagery can be enhanced and visualized to highlight specific features or spectral bands of interest.

This can include techniques such as contrast stretching, color compositing, and histogram equalization to improve image interpretation.

These preprocessing steps are essential for ensuring that Landsat imagery is accurate, calibrated, and ready for further analysis and interpretation within ArcGIS or other GIS platforms.

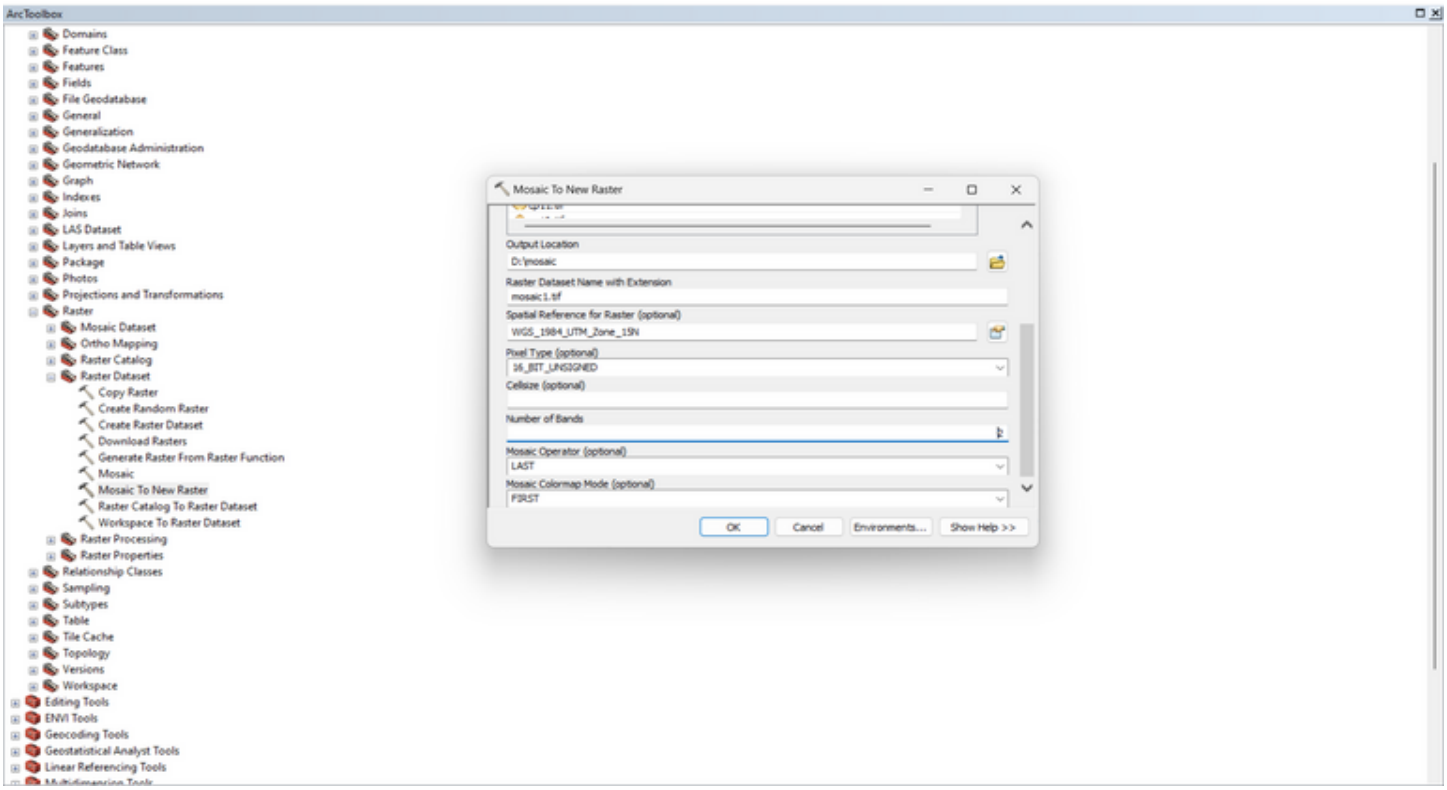


figure 11: mosaicking ArcGIS

3.2 feature engineering: beyond preprocessing

After atmospheric correction and preprocessing of remote sensing data, several advanced analysis techniques can be applied to derive valuable information for various applications.

Lineament Analysis:

Lineament analysis involves identifying linear features on the Earth's surface, such as faults, fractures, and geological boundaries, from remote sensing imagery.

Techniques for lineament analysis include visual interpretation, edge detection algorithms, and spatial analysis tools.

Lineament maps can provide insights into geological structures, tectonic activities, groundwater flow patterns, and potential areas for mineral exploration or natural hazard assessment.

Creating Band Ratios and Feature Engineering:

Band ratios involve dividing the values of one spectral band by another to enhance specific features or phenomena.

Common band ratios include Normalized Difference Vegetation Index (NDVI), which highlights vegetation health, and Normalized Difference Water Index (NDWI), which highlights water bodies.

Feature engineering involves creating new spectral indices or combinations of bands to capture specific information relevant to the study objectives.

For example, combining different bands to enhance land cover discrimination, soil moisture estimation, or urban heat island detection.

Calculating Indices like NDVI and Mineral Indices:

NDVI is a widely used index calculated from near-infrared (NIR) and red bands, given by (NIR – Red) / (NIR + Red). It provides information about vegetation density, health, and coverage.

Other indices include Normalized Difference Water Index (NDWI), Soil Adjusted Vegetation Index (SAVI), Enhanced Vegetation Index (EVI), and many more, each designed to capture specific vegetation or environmental characteristics.

Mineral indices are calculated to identify and map mineral composition or alterations in geological formations. These indices are based on the unique spectral signatures of minerals in certain wavelength regions.

Examples of mineral indices include the Normalized Difference Iron Oxide Index (NIR1 – SWIR1) / (NIR1 + SWIR1), which is sensitive to iron-bearing minerals, and the Clay Index, which helps in mapping clay minerals.

Now, for better accuracy and to detect any mistakes, we performed the NDVI calculation using two methods: one with ArcMap and the other with a Python script using the Rasterio library.

Both calculations were done before and after the atmospheric correction, and the result was very satisfying in the range between -1 and 1.

By applying these advanced analysis techniques, researchers can extract valuable information from remote sensing data for various applications such as land use and land cover mapping, vegetation monitoring, hydrological studies, geological mapping, environmental monitoring, and natural resource management. These analyses contribute to a better understanding of Earth's surface dynamics and facilitate informed decision-making processes.

snip of the python script we used for this task

```
!pip install rasterio
!pip install rioxarray
!pip install geopandas
!pip install earthpy
from google.colab import drive
import rasterio
from rasterio.windows import Window
from rasterio.enums import Resampling
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import cv2
import numpy as np
```

```
def calculate_ndvi(band_nir, band_red):
    # NDVI formula
    ndvi = (band_nir - band_red) / (band_nir + band_red)
    return ndvi

band_red = band4.read(1)
band_nir = band5.read(1)

# Calculate NDVI
ndvi = calculate_ndvi(band_nir, band_red)

# Disply abyeth w ak7el
plt.imshow(ndvi, cmap='gray', vmin=-1, vmax=1)  #  grayscale colormap
plt.colorbar(label='NDVI')
plt.title('Normalized Difference Vegetation Index (NDVI) in Black and White')
```

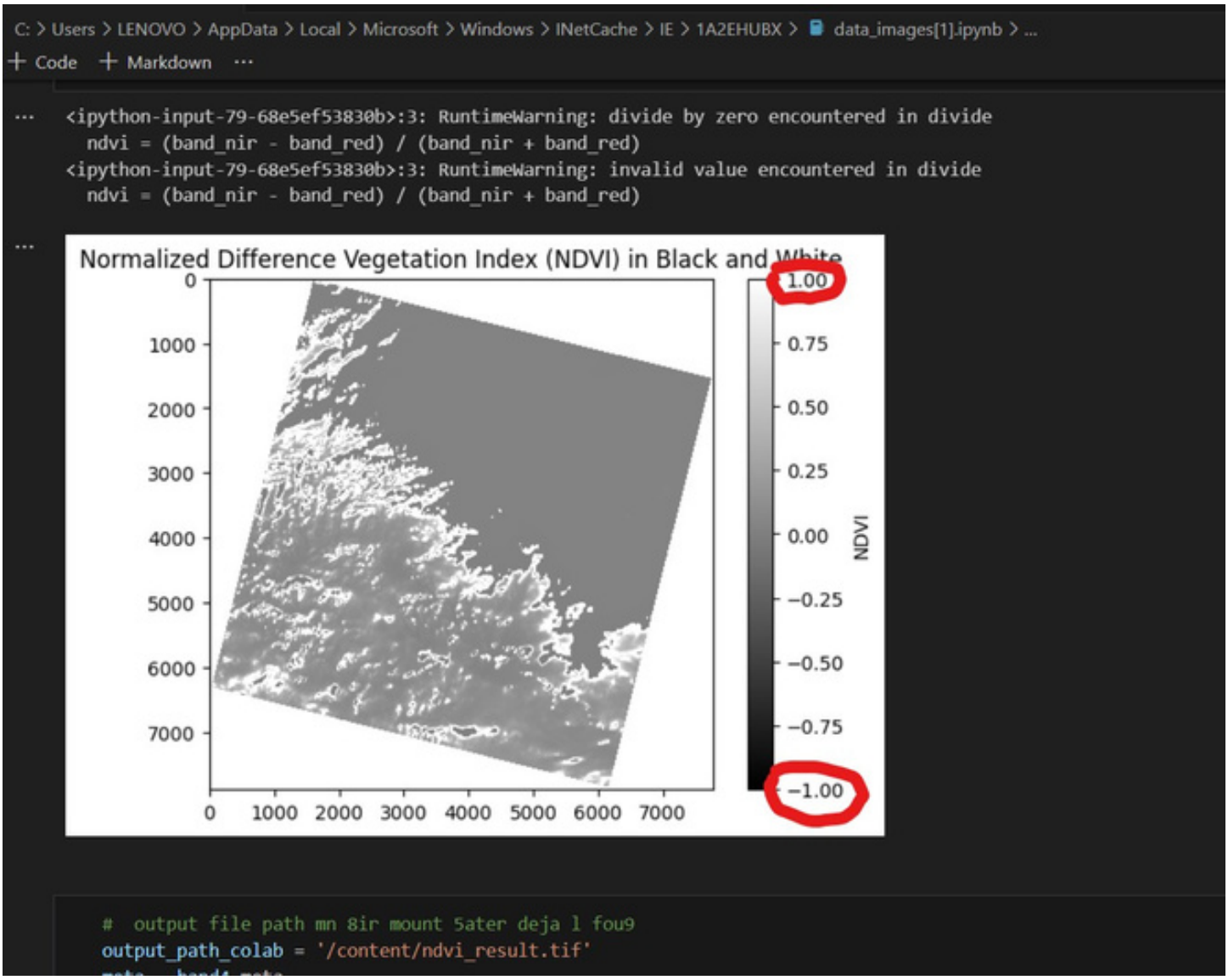


figure 12: ndvi result using rasterio python libraire

3.1.2 well logs and LAS file :

This was the most important piece of data for the project because it required expertise in the field of petroleum engineering and a good knowledge of geophysics. However, as a data scientist, facing challenges is the fun part. It pushes you beyond your limits to figure out a way to handle such delicate information and study more about the field.

Digging deeper into the University of Kansas website, we managed to retrieve the well log and the LAS file. However, the data was in a raw format and needed a lot of cleaning and preparation to make it ready for the algorithm. Thank goodness for the Python libraries. It took just a few lines of code, and the data was neat and clean.

Digging deeper into the University of Kansas website, we managed to retrieve the well log and the LAS file. However, the data was in a raw format and needed a lot of cleaning and preparation to make it ready for the algorithm. Thank goodness for the Python libraries. It took just a few lines of code, and the data was neat and clean.

3.1.2-a Acquiring the Data:

Well logs provide in-depth information about subsurface formations encountered during well drilling, often requiring expertise in petroleum engineering and geophysics for interpretation. LAS files, generated by LiDAR surveys, offer detailed 3D representations of the surface topography

The three main components formed our data. The first file was the LAS in a netcdf format containing an industry-standard file format used in all oil-and-gas and water well industries to log and store well log information and data. A

single LAS file can only contain data for one well. But in that one well, it can contain any number of datasets (called curves).

Opening this kind of file requires specific software, like Magmap, and with the features of Armap, we are able to convert the data into a suitable format. then we converted to a csv format

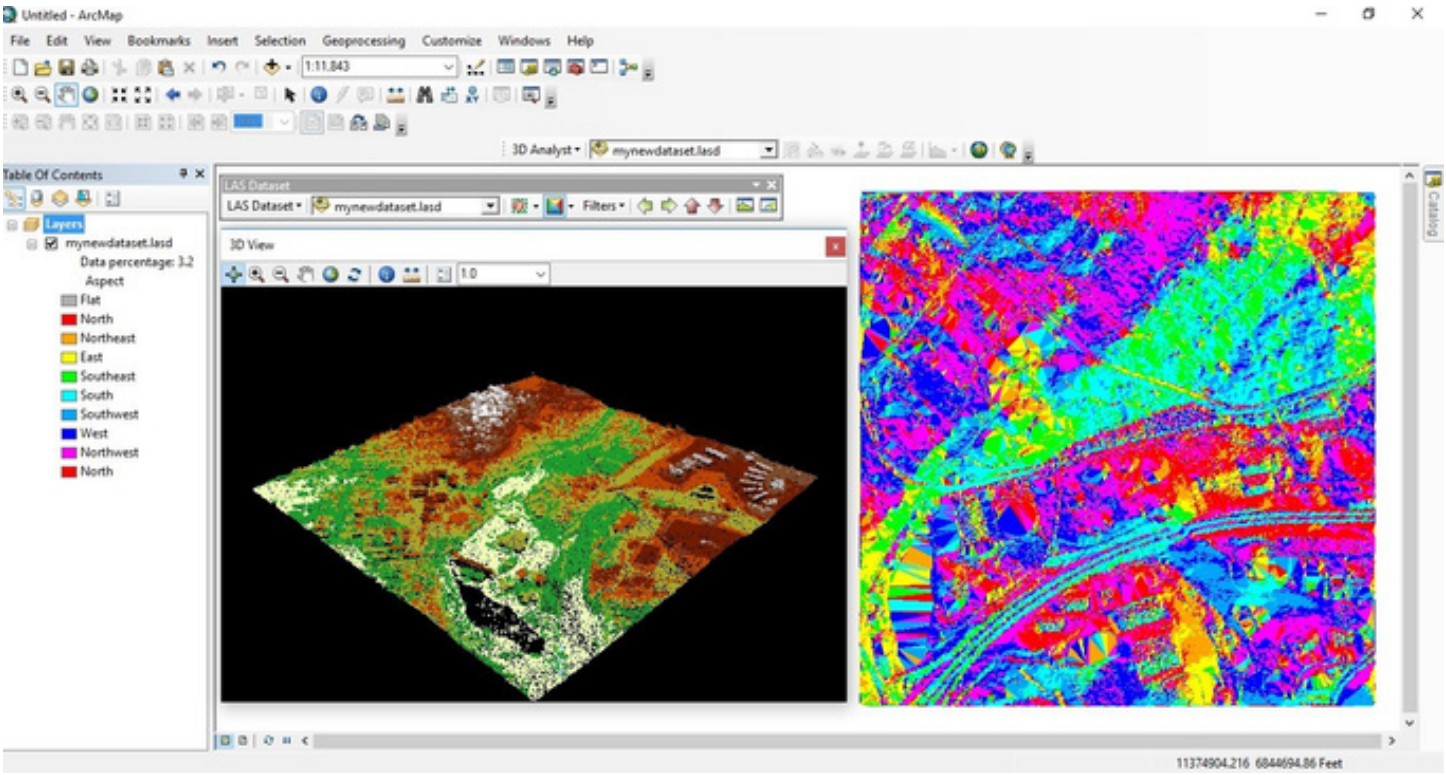


figure 13: displaying magnetic data and lidar in ArcGIS

The second one was the well logs file which provided valuable data about subsurface formations. A well log file typically contains various measurements and observations recorded during the drilling process. These logs can include information on the lithology, porosity, permeability, fluid saturation, and other properties of the formations penetrated by the well.

1. **Gamma Ray Log**: Measures natural radioactivity emitted by formations, aiding in lithology identification and correlation.
2. **Spontaneous Potential (SP) Log**: Records natural electrical potential differences between formations and drilling fluid, offering insights into fluid content and formation boundaries.
3. **Resistivity Logs**: Provide data on the electrical resistivity of formations, helping to assess fluid saturation and identify hydrocarbon-bearing zones.
4. **Density Log**: Measures the bulk density of formations, aiding in porosity determination and lithology identification.
5. **Neutron Porosity Log**: Measures the hydrogen content of formations, assisting in porosity determination and fluid identification.
6. **Sonic Log**: Records the travel time of sound waves through formations, providing data for determining formation porosity, rock mechanical properties, and depth correlation.
7. **Caliper Log**: Measures the diameter of the wellbore, aiding in assessing hole conditions and well integrity.
8. The last part was simple yet effective. First, we took the geologic map of Kansas City and extracted the information from the legend of the map, which helped us convert this information into a CSV table.

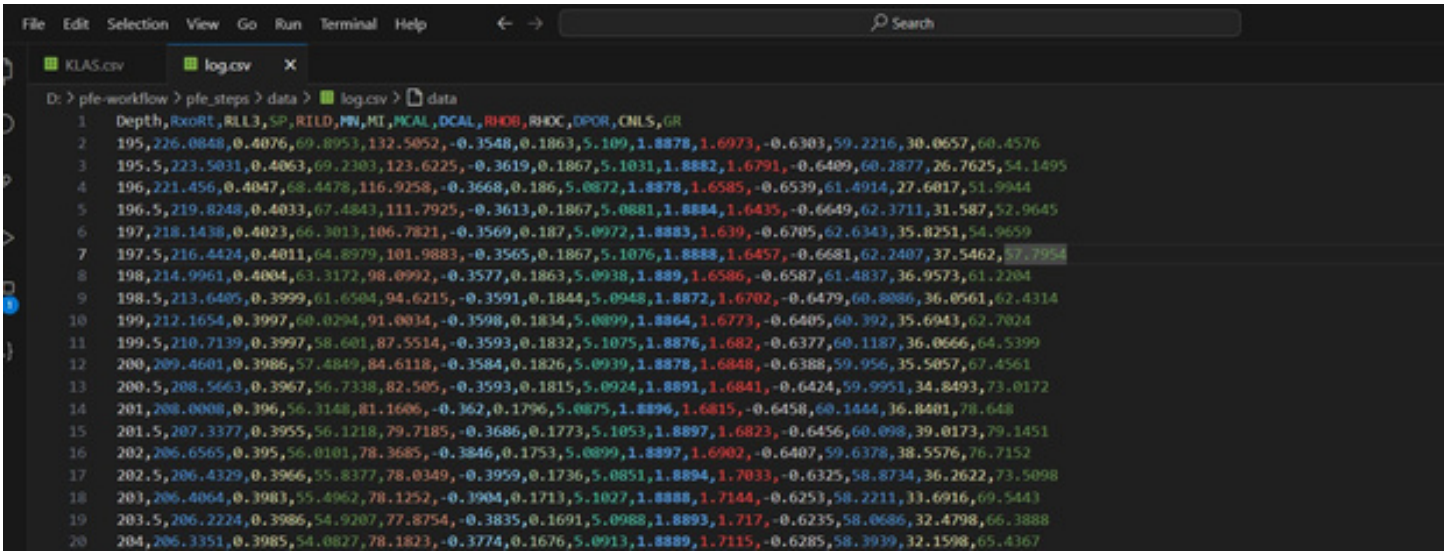


figure 14: displaying well log file in vscode

Aeromagnetic surveys are able to identify changes in the Earth's magnetic field that result from various geological structures and types of rock. Geologists utilize aeromagnetic data in oil and gas exploration to locate faults, folds, and anticlines, among other structural characteristics.It was about time to use a basic Python script to process the magnetic data even more. Finding the area's geological faults was the major objective.

```
import cv2

import numpy as np

import pandas as pd

import hemi_fault_detection as hf

img = cv2.imread('ln_og.png')

from IPython.display import Image

for points in lines:

    x1,y1,x2,y2 = points[0]

    cv2.line(img, (x1,y1),(x2,y2),(0,0,0))

cv2.imwrite('faults.png', img)
```

```
Image('faults.png')
```

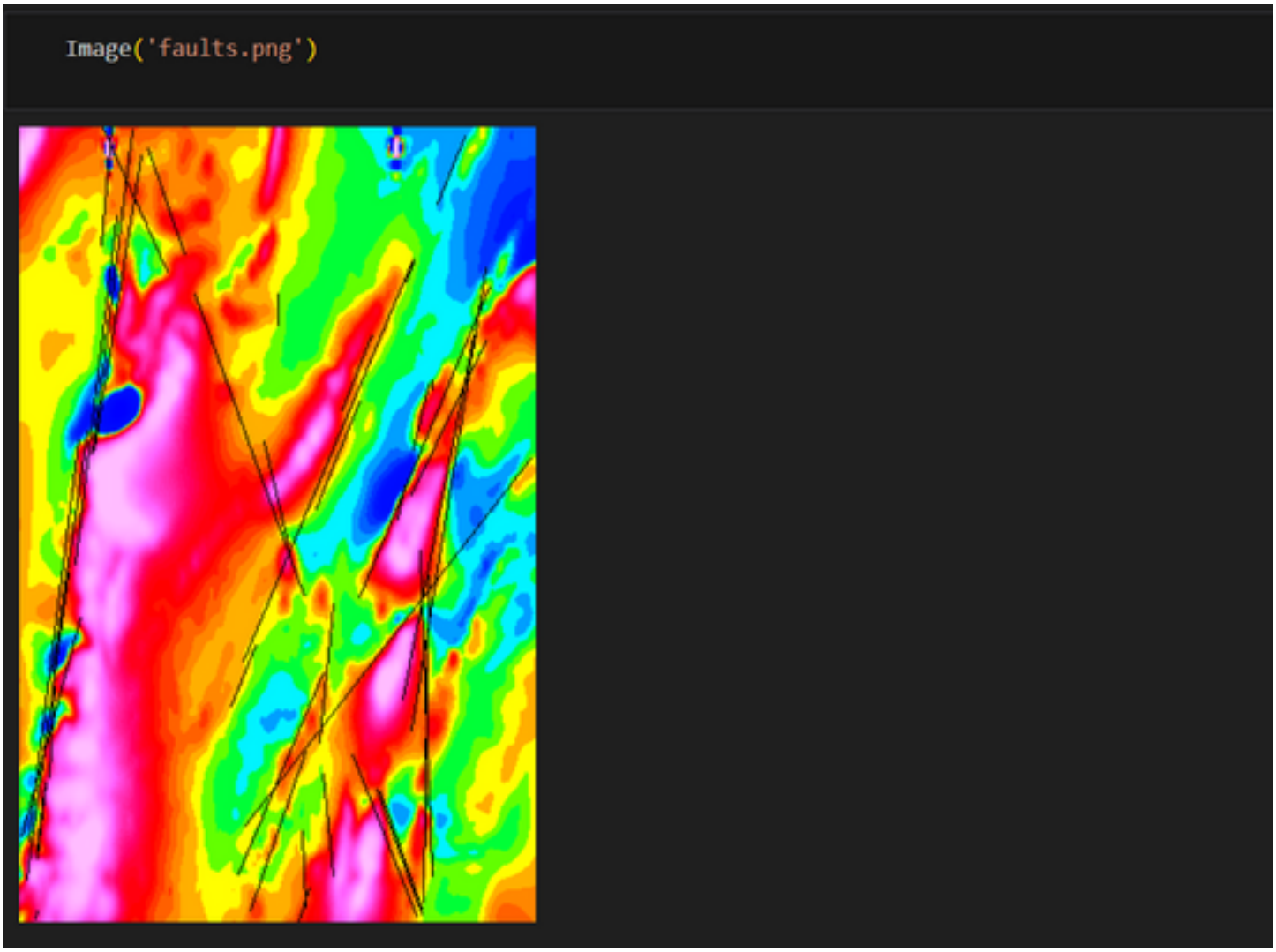


figure 15: result of extracting fracture in the image

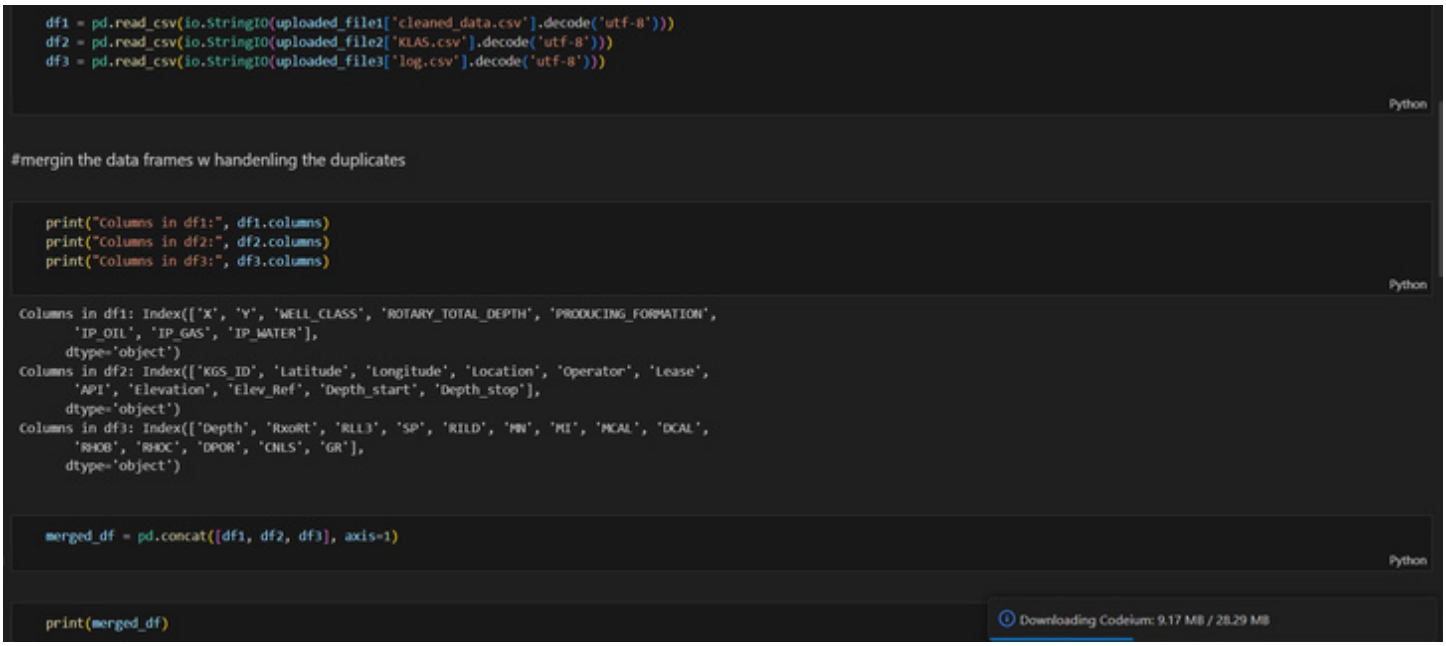
For most cases, the data was already there on Kuggle or any other platform for data science, but in our case, we created the data on our own since we have all the files we need. Now we're going to need to perform a combination to create our dataset. which takes us to the next part of this section.

3.1.2-b Data Preprocessing with Python Libraries:

Leveraging Python libraries specifically designed for geoscience data manipulation, we streamlined the data cleaning and preparation process. These libraries facilitated efficient handling of the well logs and LAS file data, transforming them into a structured and clean format suitable for our model's analysis

First, we needed to get our priority straight, which is transforming all the files into a suitable format for Python to understand, and this format was a csv file. Then, we performed a combination and merged the files into one.

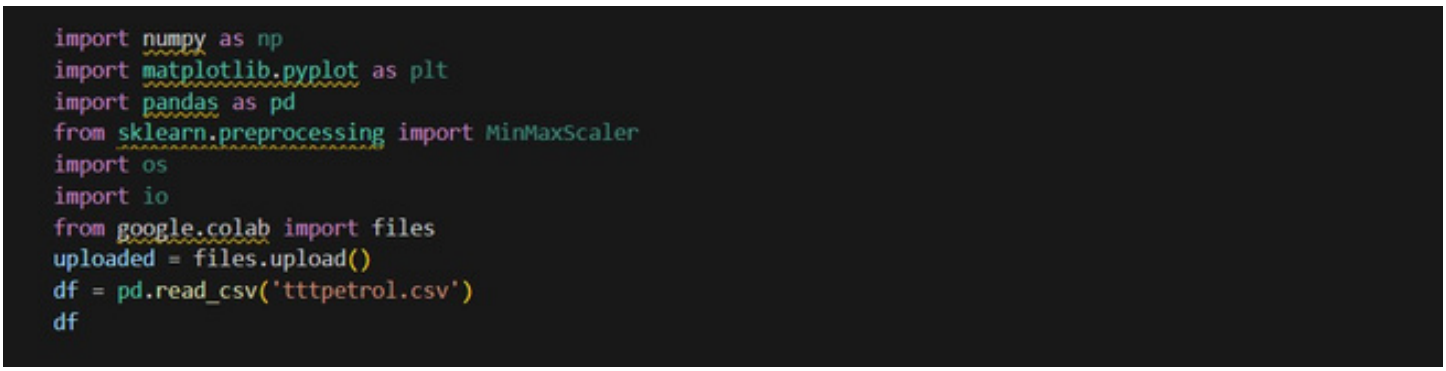
Once we have our file ready, we look deep into the data and try to understand the concepts in it to know what is essential and what is irrelevant for our project**.**



While the first step is done, it's time for the real work to start. Dividing it into small tasks was my go-to to get it done. In this paragraph, we will go over these steps, explaining step by step how we managed to come up with the dataset we used later for the project.

Breakdown of the Data Preprocessing Code for Well Log Data This code performs several essential data preprocessing steps **1-Importing Libraries and Loading Data:**

Necessary libraries for data manipulation (pandas), visualization (matplotlib), and scikit-learn for scaling are imported (numpy, matplotlib.pyplot, pandas, scikit-learn, os, io)



2-Data Cleaning: Handling Missing Values:

The code removes leading/trailing spaces from column names using df.columns = df.columns.str.strip().

It then checks for missing values by finding the sum of null values in each column and storing the result in missing values prints the columns with missing values,

revealing which columns have significant data absence

```
df.columns = df.columns.str.strip()
# Check for missing values fil colums
missing_values = df.isnull().sum()
print("Columns with missing values:")
print(missing_values[missing_values > 0])
```

```
Columns with missing values:
WELL_CLASS              5946
ROTARY_TOTAL_DEPTH     10663
PRODUCING_FORMATION    55998
IP_OIL                 64331
IP_GAS                 67950
IP_WATER               66878
KGS_ID                 61701
```

3- identifying numerical and categorical columns

separating the DataFrame columns into two categories: numerical (numerical_cols) and categorical (categorical_cols) based on their data types (integer, float, or object).

```
missing_values = {
    'WELL_CLASS': 5946,
    'ROTARY_TOTAL_DEPTH': 10663,
    'PRODUCING_FORMATION': 55998,
    'IP_OIL': 64331,
    'IP_GAS': 67950,
    'IP_WATER': 66878,
    'KGS_ID': 61701,
    'Latitude': 61701,
    'Longitude': 61701,
    'Location': 61701,
}
# Inspect
print("Columns in DataFrame:")
print(df.columns)
# Identify numerical and categorical columns
numerical_cols = [col for col in df.columns if df[col].dtype in ['int64', 'float64']]
categorical_cols = [col for col in df.columns if df[col].dtype == 'object']
# Compare withe data l 'origin
missing_numerical_cols = [col for col in numerical_cols if col in missing_values.keys()]
missing_categorical_cols = [col for col in categorical_cols if col in missing_values.keys()]
print("Numerical columns with missing values:", missing_numerical_cols)
print("Categorical columns with missing values:", missing_categorical_cols)
numerical_cols.extend(missing_values.keys())
# Include additional categorical columns
categorical_cols.extend(missing_values.keys())
# Fill missing values in numerical columns with mean
for col in numerical_cols:
    if col in missing_values.keys():
        if df[col].dtype != 'object':
            df[col].fillna(df[col].astype(float).mean(), inplace=True)
        else:
            print(f"Skipping column '{col}' due to non-numeric values")
```

4-Targeting Missing Values in Different Data Types:

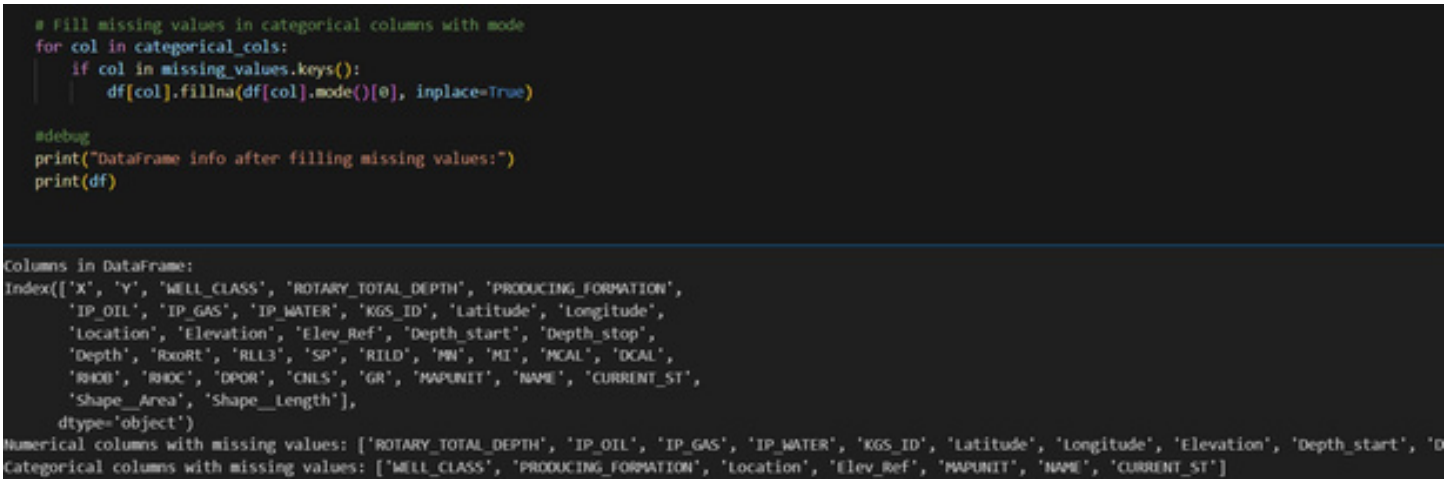
It identifies missing values specifically in numerical and categorical columns using list comprehensions.

The code employs different strategies to address missing values based on data type:

Numerical Columns: The mean value of each column (excluding missing values) is calculated and used to fill in the missing entries using df[col].fillna(df[col].astype(float).mean(),inplace=True).

Columns containing non-numeric values are skipped.

Categorical Columns: The most frequent value (mode) within each column is used to fill in the missing entries using df[col].fillna(df[col].mode()[0], inplace=True).



```
# Fill missing values in categorical columns with mode
for col in categorical_cols:
    if col in missing_values.keys():
        df[col].fillna(df[col].mode()[0], inplace=True)

#debug
print("DataFrame info after filling missing values:")
print(df)
```

```
Columns in DataFrame:
Index(['X', 'Y', 'WELL_CLASS', 'ROTARY_TOTAL_DEPTH', 'PRODUCING_FORMATION',
       'IP_OIL', 'IP_GAS', 'IP_WATER', 'KGS_ID', 'Latitude', 'Longitude',
       'Location', 'Elevation', 'Elev_Ref', 'Depth_start', 'Depth_stop',
       'Depth', 'RxoRt', 'RLL3', 'SP', 'RILD', 'MN', 'MI', 'MCAL', 'DCAL',
       'RHOB', 'RHOC', 'DPOR', 'CNLS', 'GR', 'MAPUNIT', 'NAME', 'CURRENT_ST',
       'Shape__Area', 'Shape__Length'],
      dtype='object')
Numerical columns with missing values: ['ROTARY_TOTAL_DEPTH', 'IP_OIL', 'IP_GAS', 'IP_WATER', 'KGS_ID', 'Latitude', 'Longitude', 'Elevation', 'Depth_start', 'D
Categorical columns with missing values: ['WELL_CLASS', 'PRODUCING_FORMATION', 'Location', 'Elev_Ref', 'MAPUNIT', 'NAME', 'CURRENT_ST']
```

5-handeling duplicate rows:

. A new Data Frame df1 is created as a copy of the updated data (df.copy()).

. The code identifies and prints the number of duplicate rows present in df1 using df1[df1.duplicated().

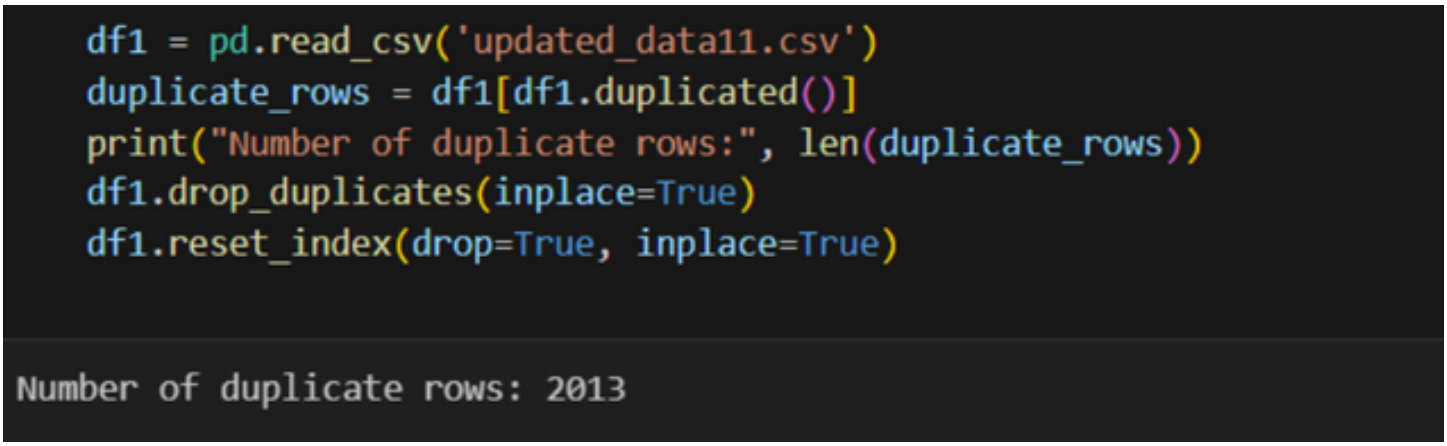
. It then removes duplicate rows from df1 using df1.drop_duplicates(inplace=True).

. The Data Frame without duplicates is reset to ensure proper indexing using df1.reset_index(drop=True,

inplace=True).

. Finally, the data without duplicates is saved as a new CSV file named

"updated_data_with_no_duplicates.csv" and offered for download.



```
df1 = pd.read_csv('updated_data11.csv')
duplicate_rows = df1[df1.duplicated()]
print("Number of duplicate rows:", len(duplicate_rows))
df1.drop_duplicates(inplace=True)
df1.reset_index(drop=True, inplace=True)
```

```
Number of duplicate rows: 2013
```

1. **Data normalization:**

Another copy of the data (df2) is created from df1.

A list named columns_to_normalize specifies the columns that will undergo normalization. These columns likely represent continuous features that might benefit from scaling to a common range between 0 and 1.

A MinMaxScaler object is created from scikit-learn.

The code applies the MinMaxScaler to the specified columns in df2 using df2[columns_to_normalize] = scaler.fit_transform(df1[columns_to_normalize]). This normalizes the values within these columns.

The normalized Data Frame (df2) is printed forinspection. Finally,the normalized data is saved as a new CSV file named "normalized.csv" and offered for download

```
from sklearn.preprocessing import MinMaxScaler
df2 = df1.copy()
columns_to_normalize = ['X', 'Y', 'ROTARY_TOTAL_DEPTH', 'Elevation', 'Depth_start', 'Depth_stop', 'Depth',
                        'RxoRt', 'RLL3', 'SP', 'RILD', 'MN', 'MI', 'MCAL', 'DCAL', 'RHOB', 'RHOC',
                        'DPOR', 'CNLS', 'GR', 'Shape__Area', 'Shape__Length']
scaler = MinMaxScaler()
df2[columns_to_normalize] = scaler.fit_transform(df1[columns_to_normalize])
print(df2)
```

```
          X         Y            WELL_CLASS  ROTARY_TOTAL_DEPTH  \
0  0.680132  0.085776  Plugged and Abandoned            0.300073
1  0.581537  0.359137  Plugged and Abandoned            0.396659
```

Conclusion :

Overall, we effectively demonstrate a data preprocessing workflow for the combined data, including handling missing values, identifying data types, applying appropriate techniques for each type, removing duplicates, and normalizing specific columns. which is essential for preparing the dataset for the machine learning model. Our data integration process involved merging well logs, LAS files, remote sensing data, and geologic information into a single comprehensive dataset using Python's Pandas library.

We addressed inconsistencies in data formats and units and performed feature engineering to extract relevant geological features from the integrated dataset

Chapter 5: Conducting the AI/QML Symphony in Resource Exploration

prelude:

We have come this far with the project. and the real challenge begins. Pushing through our limits was a requirement here since my field is geomatics, and getting this far into the world of binary numbers and going even beyond that, In this chapter, we will explore the fascinating world of artificial intelligence (AI) and applying quantum physics to computing We'll talk about AI, QML, and how tech impacts energy. These ideas are key. We use a mixed way to conquer our work. The sentences vary: short, long, simple, complex.

1-concept:

In the current age of the Fourth Industrial Revolution, the digital world has a wealth of data, such as Internet of Things data, cybersecurity data, health data, etc. To intelligently analyze these data and develop the

corresponding *smart and automated* applications, knowledge of artificial intelligence (AI), particularly *machine learning (ML),* is the key. Various types of machine learning algorithms, such as supervised, unsupervised, semisupervised, and reinforcement learning, exist in the area. Besides, *deep learning*, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale.In this chapter, we present a comprehensive view of these algorithms that can be applied to enhance the intelligence and capabilities of an application. Thus, this section's key contribution is explaining the principles of different techniques and their applicability in various real-world application domains. We also highlight the challenges and potential research directions based on our study. Overall, these nexus lines aim to serve as a reference point. in the context of data analysis and computing that typically allow the applications to function in an intelligent manner. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular and latest technology in the fourth industrial revolution. The learning algorithms can be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning.

1-1 Types of Real-World Data :

Structured: It has a well-defined structure, conforms to a data model following a standard order that is highly organized and easily accessed, and is used by an entity or a computer program.

Unstructured: On the other hand, there is no pre-defined format or organization for unstructured data, making it much more difficult to capture, process, and analyze. It mostly contains text and multimedia material. For example, sensor data

Semi-structured: Semi-structured data is not stored in a relational database like the structured data mentioned above, but it does have certain organizational properties that make it easier to analyze. HTML, XML, JSON documents, NoSQL databases, etc.,

Metadata: It is not the normal form of data, but “data about data." The primary difference between “data” and “metadata” is that data is simply the material that can classify, measure, or even document something relative to an organization’s data properties. On the other hand, metadata describes the relevant data information, giving it more significance for data users. A basic example of a document’s metadata might be the MTL file that comes with the Landsat bands containing the sun's elevation and reflectance values.

1-2 Machine Learning Techniques:

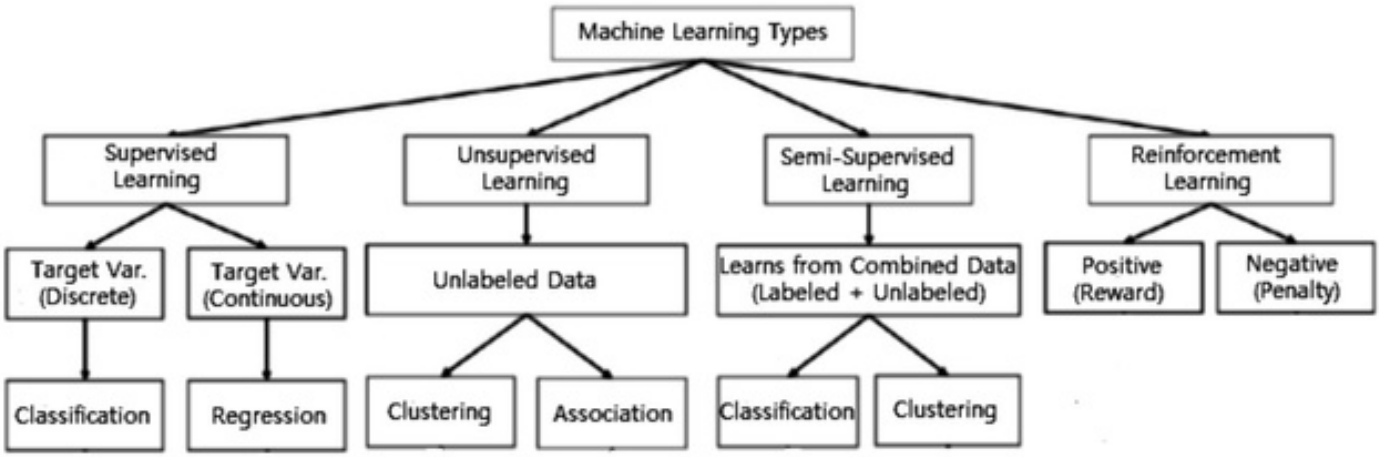


figure 15: Various types of machine learning techniques

Supervised learning: involves machine learning tasks where a function is learned to map inputs to outputs based on labeled input-output pairs. It utilizes labeled training data and aims to infer a function using a collection of training examples. Supervised learning is task-driven, with common tasks including classification to separate data and regression to fit the data. An example of supervised learning is text classification, where the goal is to predict the class label or sentiment of a piece of text, such as a tweet or a product review.

Unsupervised learning: involves analyzing unlabeled datasets without human interference. It is used to extract generative features, identify trends and structures, group results, and for exploratory purposes. Common tasks in unsupervised learning include clustering, density estimation, feature learning, dimensionality reduction, finding association rules, and anomaly detection.

Semi-supervised: learning combines supervised and unsupervised methods by operating on both labeled and unlabeled data. It is useful in contexts where labeled data is rare and unlabeled data are abundant. The goal of semi-supervised learning is to improve prediction outcomes compared to using labeled data alone.

Applications of semi-supervised learning include machine translation, fraud detection, labeling data, and text classification.

Reinforcement learning: is a type of machine learning algorithm that enables software agents and machines to automatically evaluate optimal behavior in a specific context or environment to improve efficiency. This learning method is environment-driven and relies on rewards or penalties. Its goal is to use insights from the environment to maximize rewards or minimize risks. Reinforcement learning is useful for training AI models to increase automation or optimize the operational efficiency of complex systems such as robotics, autonomous driving, manufacturing, and supply chain logistics. However, it is not preferred for solving basic or straightforward problems.

Many classification algorithms have been proposed in the machine learning and data science literature. In the following, we summarize the most common and popular methods that are widely used in various application areas.

1-3 Machine Learning algorithms:

Linear Discriminant Analysis (LDA) : is a linear decision boundary classifier that fits class conditional densities to data and applies Bayes' rule. It is a generalization of Fisher's linear discriminant, reducing the dimensionality of a dataset to minimize model complexity and computational costs. The standard LDA model assumes Gaussian density for each class and a shared covariance matrix. LDA is closely related to ANOVA and regression analysis, aiming to express a dependent variable as a linear combination of other features or measurements

Logistic Regression (LR) : is a probabilistic statistical model commonly used for classification in machine learning. It employs a logistic function to estimate probabilities and can overfit high-dimensional datasets. Regularization techniques such as L1 and L2 can be utilized to prevent overfitting. However, the assumption of linearity between dependent and independent variables is considered a major drawback. LR can be applied to both classification and regression problems, but it is predominantly used for classification.

$$g(z) = \frac{1}{1 + \exp(-z)}.$$

K-Nearest Neighbors (KNN): is an instance-based learning algorithm that stores all training data instances in n-dimensional space and classifies new data points based on similarity measures such as the Euclidean distance function. It uses a majority vote of the k nearest neighbors to compute classification. KNN is robust to noisy training data, and its accuracy depends on data quality. However, selecting the optimal number of neighbors is a significant challenge. KNN can be applied to both classification and regression tasks.

Decision Tree (DT) : is a non-parametric supervised learning method used for both classification and regression tasks. It includes well-known algorithms such as ID3, C4.5, and CART, as well as more recent ones like BehavDT and IntrudTree, which are effective in specific application domains. DT classifies instances by sorting down the tree from the root to leaf nodes, checking the attribute defined by each node. The most popular criteria for splitting are "gini" for Gini impurity and "entropy" for information gain.

Entropy : H(x)=− (2) ∑n i=1 p(xi)log2 p(xi)

Gini(E) =1 − (3) ∑c i=1 pi 2

1-4 Dimensionality Reduction and Feature Learning

The two main techniques for dimensionality reduction in machine learning and data science are feature selection and feature extraction.

1. **Feature Selection:**

Involves retaining a subset of the original features.

Aims to select the most relevant features for the problem at hand.

Helps in reducing the dimensionality of the data while retaining its interpretability.

1. **Feature Extraction:**

Involves creating new features based on the original ones.

Aims to combine the existing features to form a smaller set of new features.

Can help in capturing the most important information in the original features while reducing dimensionality.

Both of these techniques are used to address the challenges presented by high-dimensional data in machine learning and data science.

1-5Deep Learning in Resource Exploration:

1-5.1 intro:

While machine learning focuses on algorithms and models to parse data, learn from it, and make informed decisions, deep learning takes this a step further by mimicking the human brain to process data and create patterns for use in decision making. This approach, inspired by the structure and function of the brain's neural networks, has proven to be particularly effective in tasks such as image and speech recognition, natural language processing, and more. The shift to deep

learning opens up a world of possibilities for solving complex problems and unlocking new frontiers in artificial intelligence.

This new technology can actually be categorized into two main sections: artificial neural networks (ANN) and convolutional neural networks (CNN).

1-5.2 Artificial Neural Networks (ANN)

Well, we can find a hundred definitions for neural networks on Google, but by a simple definition, I would like to call it a simulation of the human brain mechanism. Our brains identify things through certain patterns. And according to these patterns, the brain activates certain neurons to trigger specific areas of the brain, like releasing cortisol and adrenaline, which alerts the nervous system. stimulated by the amygdala. the same principles applied to computers: by feeding it a certain input,neural networks process data by recognizing patterns and adjusting the connections between artificial neurons (nodes) targeted neurons get activated, doing the calculations of the weight and the bios to get the wanted output.

1-5.3architecture

The neural network architecture encompasses interconnected layers of neurons, categorized into input, hidden, and output layers. Data is initially received by the input layer and then undergoes processing through multiple hidden layers via weighted connections and activation functions. Subsequently, the output layer generates the network's final predictions or classifications. Each connection between neurons possesses an associated weight, which is modified during training to enhance the network's predictive abilities. The network learns by iteratively adjusting these weights based on prediction errors, facilitated through a process known as backpropagation. This brain-inspired architecture empowers neural networks to effectively address a wide array of tasks, spanning image and speech recognition to natural language processing. The neural network structure comprises neurons, layers, and weights, where neurons are organized into layers, and weights are utilized to regulate the influence of various inputs and biases. The intricacy of this architecture's design is intricately linked to the specific task being addressed

In a Neural Network, the flow of information occurs in two ways –

Feedforward Networks: In this model, the signals only travel in one direction, towards the output layer.

Feedforward Networks have an input layer and a single output layer with zero or multiple hidden layers.

They are widely used in pattern recognition.

Feedback Networks: In this model, the recurrent or interactive networks use their internal state (memory) to process the sequence of inputs. In them, signals can travel in both directions through the loops (hidden layer/s) in the network. They are typically used in time-series and sequential tasks

1-5.4 Components

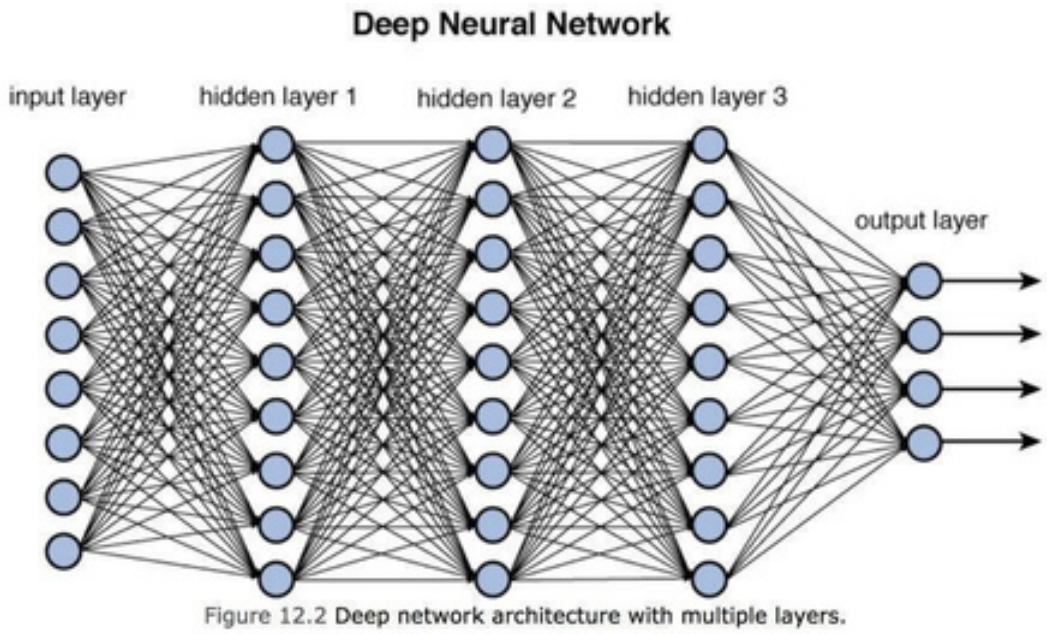


figure 21: deep neural network architecture

A neuron, also known as an activation function, serves as the fundamental unit of a neural network, receiving input from external sources or other nodes. Each node establishes connections with nodes in the subsequent layer, with each connection possessing a specific weight reflective of the neuron's relative importance against other inputs. The input layer's node values, multiplied by their respective weights and aggregated, produce a value for the first hidden layer. Subsequently, the hidden layer features a pre-defined "activation" function that determines the node's activation and intensity based on the aggregated value. This hidden layer, concealed from the external domain, undertakes the primary computational tasks within the neural network, processing inputs from the input layer to generate results forwarded to the output layer for user accessibility.

Consider an example: an image depicting an apple, composed of 28 pixels by 28 pixels In this scenario, the output value ranging from 0 to 1 serves as an indicator. If the value falls below 0.5, it identifies the object as an orange; conversely, if it surpasses 0.5, it recognizes it as an apple. As we move to the hidden layer, the network discerns patterns primarily based on shape rather than color. As the input data traverses through the network, different neurons become activated, a process termed forward propagation.

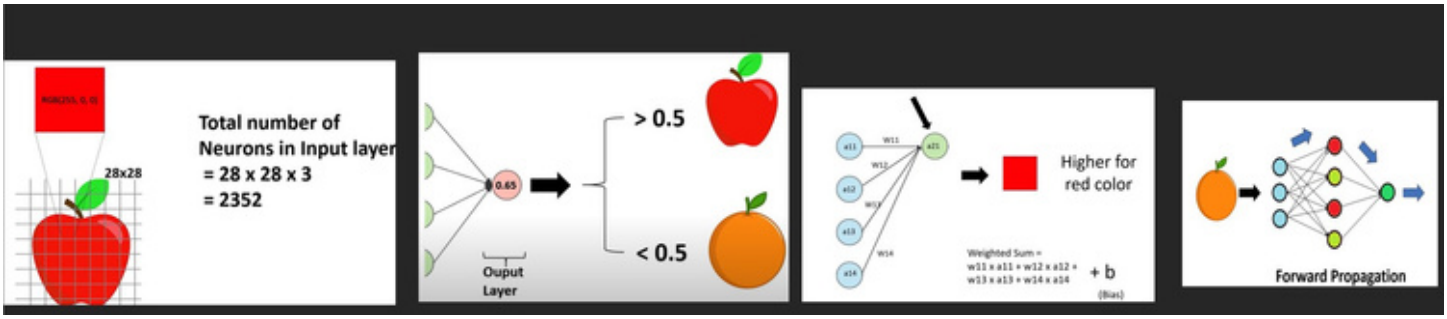


figure 22: mechanism explained

1-5.5 algorithms:

The learning (or training) process is initiated by dividing the data into three different sets:

Training dataset: This dataset allows the neural network to understand the weights between nodes.

Validation dataset: This dataset is used for fine-tuning the performance of the neural network.

Test dataset: This dataset is used to determine the accuracy and margin of error of the neural network.

The process of training a Neural Network involves segmenting data into training, validation, and testing sets. Optimization algorithms are then applied to facilitate training, adjusting the network's parameters to minimize errors. These algorithms vary in characteristics such as memory requirements, numerical precision, and processing speed, each offering unique advantages and considerations

What is the Learning Problem?

In machine learning, we frame the learning process as minimizing a loss function (f), which evaluates performance on a dataset. This function typically comprises an error term and a regularization term. While the error term assesses the fit to the data, the regularization term mitigates overfitting by managing complexity. The loss function ($f(w)$) depends on adaptive parameters (weights and biases), represented as a single ndimensional weight vector (w).

Now that we know what the learning problem is, we can discuss the five main algorithms

1. One-dimensional optimization:

Since the loss function depends on multiple parameters, one-dimensional optimization methods are instrumental in training .algorithms first compute a training direction (d) and then calculate the training rate (η) that helps minimize the loss in the training direction [$f(\eta)$]

Golden Section Method :

The golden section search algorithm is used to find the minimum or maximum of a single-variable function [f(x)]. If we already know that a function has a minimum between two points, then we can perform an iterative search just like we would in the bisection search for the root of an equation f(x) = 0

Brent's Method:

is a root-finding algorithm that combines root bracketing, bisection, secant, and inverse quadratic interpolation. It employs a Lagrange interpolating polynomial of degree 2 and is expressed as "Method: Brent in Find Root[eqn, x, x0, x1]." provided the function values are computable within a specific region containing a root. It fits a quadratic function of y to x using interpolation formula when given three points x1, x2, and x3.

2. Multidimensional optimization

we know that the loss function is a non-linear function of the parameters, it is impossible to find the closed training algorithms for the minimum. However, if we consider searching through the parameter space, which includes a series of steps, at each step, the loss will be reduced by adjusting the parameters

In this case the nn is trained by choosing a random parameter vector and then generating a sequence of parameters to ensure that the loss function decreases with each iteration of the algorithm.

Here are three examples of multidimensional optimization algorithms:

Gradient descent:

The gradient descent algorithm, a simple training method, uses information from the gradient vector, making it a first-order method. It starts at w(0) and progresses iteratively until a specified criterion is met. At each step, it updates the weight w(i) by moving in the opposite direction of the gradient vector g(i) multiplied by a learning rate $\eta(i)$. The iteration follows the formula:

$w(i+1) = w(i) - g(i)\eta(i)$

Newton's method:

Newton's method is a second-order algorithm utilizing the Hessian matrix. It improves training directions by incorporating second derivatives of the loss function. Denoting f[w(i)] as f(i), ∇ f[w(i)] as g(i), and Hf[w(i)] as H(i), we consider the quadratic approximation of f at w(0) using Taylor's series expansion:

a. = f(0) + g(0) · [w−w(0)] + 0.5 · [w−w(0)]^2 · H(0)

Here, H(0) is the Hessian matrix of f at w(0). Setting g = 0 for the minimum of f(w), we obtain:

a. = g(0)+H(0) · (w−w(0)) = 0

Thus, starting from the parameter vector w(0), Newton's method iterates as:

w(i+1) = w(i)−H(i)^−1 · g(i)