pfe project

Chapter 4: Study conception from pixels to predictions

Introduction :

in This chapter we are gonna dig deeper into the roadmap of our AI/QML-driven resource exploration project: the study conception. 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. This includes the selection of the AI/QML development platform and model, the data acquisition process for various sources (satellite imagery, geospatial data, fieldwork data), and the crucial 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 AI/QML model for successful resource exploration

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 softwares 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. Many users are familiar with its interface and workflows.

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.

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 crucial for AI/QML development

The 64-bit architecture of ArcGIS Pro facilitates working with large datasets, 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 individuals 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).

Might 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 are crucial 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.

Steeper 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 magnetic anomalies and identifying potential subsurface structures.

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 can be expensive, particularly for individual users or smaller research projects with limited budgets. 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, making it a cost-effective option.

Extensive Language Support: Offers support for numerous programming languages, including Python, 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: Built-in Git support enables version control and collaboration on geospatial projects.

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: Users may need to install extensions or plugins to add GIS-related functionalities, which may not always provide the same level of integration and ease of use as dedicated GIS software.

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.

Learning Curve: While VS Code is generally user-friendly, users may need to invest some time in learning its features and customization options, especially if they are new to coding or development environments.

Conclusion:

In the realm of geospatial analysis and resource exploration, a variety of software platforms cater to different needs and project requirements. Having experience with ArcGIS, QGIS, and ENVI allows for a flexible approach depending on the specific task at hand.

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

Data Integration: Seamlessly combining remote sensing data, geospatial information, and fieldwork data is crucial for our AI/QML model development.

AI/QML Integration: ArcGIS Pro's Python 3 support simplifies incorporating AI/QML libraries for model building and analysis.

User-friendliness: If your team is familiar with GIS concepts, ArcGIS offers a balance between advanced functionalities and a user-friendly interface.

QGIS offers a valuable complementary tool, particularly for its:

Cost-Effectiveness: The open-source nature of QGIS makes it a great option for exploring custom workflows or initial data visualizations.

Python Scripting: Similar to ArcGIS Pro, QGIS supports Python scripting, allowing for further customization and potential integration with AI/QML libraries (though requiring more technical expertise).

ENVI can be a valuable asset when needed, providing advanced image processing capabilities for tasks like:

Atmospheric Correction: Correcting for atmospheric effects is crucial for accurate interpretation of remote sensing data in resource exploration.

Spectral Band Manipulation: ENVI's functionalities for manipulating spectral bands can be instrumental in feature extraction and preparing data for AI/QML models.

MAGMAP offers advanced magnetic data processing and interpretation tailored for geophysical exploration tasks, making it a valuable addition to the toolkit when analyzing magnetic anomalies and subsurface structures.

Visual Studio Code serves as a versatile development environment, facilitating scripting, customization, and integration with other tools commonly used in GIS workflows.

By leveraging the strengths of each software platform throughout the project, we can ensure a comprehensive workflow for AI/QML-driven resource exploration, from data acquisition and processing to model development and analysis.

This revised conclusion acknowledges the value of all six software's options while highlighting ArcGIS Pro as our primary choice due to the project's specifics (data integration, AI/QML focus). It also showcases my expertise in working with all 6 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 the 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, focusing on selecting spectral bands that effectively detect hydrocarbons. This selection will consider the trade-off between spatial resolution (e.g., 30 meters) and data volume. The downloaded Landsat data, in a format compatible with our GIS platform (e.g., ArcGIS), will form a crucial 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

figure5 : 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 find 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 and charm..

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 cappotential to save hydrocarbons.

Permeability:

Reflects the rock's cappotential 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 logs function crucial equipment for geoscientists and engineers to recognize subsurface situations and optimize drilling and manufacturing strategies.

figure6 : 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 critical. Here's how well logs and geological data are crucial in this context:

Identification of Potential Reservoirs: Geological maps, regional seismic data, and existing studies provide valuable insights into the geological formations present in the area. By analyzing these data, geologists can identify potential hydrocarbon reservoirs based on the presence of suitable rock formations known to host hydrocarbons.

Evaluation of Rock Characteristics: Well logs offer detailed information about the characteristics of rock formations encountered during drilling. This includes data on lithology, porosity, permeability, and other parameters crucial for assessing the potential of a reservoir to store and produce hydrocarbons. Geologic information helps interpret these well logs by providing context about the depositional environment and the geological history of the area.

Delineation of Trap Structures: Certain geological features, such as anticlines, faults, and stratigraphic traps, can act as traps for hydrocarbons. By analyzing 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, aiding in the delineation of potential hydrocarbon accumulations.

Estimation of Depth and Thickness: Well logs provide data on the depth and thickness of different geological layers, including potential hydrocarbon-bearing formations. By correlating well log data with geological information, geologists can estimate the depth and thickness of potential hydrocarbon zones within the subsurface, helping in the planning and execution of drilling operations.

Validation with Historical Well Data: Historical well data from previous drilling activities in the area can provide valuable validation for geological interpretations and predictions. Well logs from existing wells allow geologists to compare and correlate geological formations across different locations, improving the accuracy of subsurface models and predictions for hydrocarbon exploration.

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. By integrating data from various sources and employing advanced geological interpretation techniques, geoscientists can make informed decisions and optimize exploration efforts for discovering and developing hydrocarbon resources.

figure 7 : 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 AI/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.

figure 8: 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. Integrating these maps with other data sources like well logs, seismic data, and geologic information further enhances our understanding of the subsurface geology and potential hydrocarbon reservoirs.

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"}
```

figure 9: step 1 of workflow

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 10: creating composite band in ArcGIS pro

figure 11: enhancing resolution in ArcGIS pro

figure 12: checking proprieties and raster info 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.

figure 13: performing geometric correction in ArcGIS pro

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.

figure 14: USGS figure of toa values and conversion

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 here we used ArcMap for more efficiency

figure 15: using raster function in ArcMap

figure 16: ortho corrected landsat 9

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.

figure 17: mosaicking ArcGIS

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.

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.

figure 18: ndvi result in ArcMap

snip of the python script we used for this task

figure 19: 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.

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 17: 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.

Gamma Ray Log: Measures natural radioactivity emitted by formations, aiding in lithology identification and correlation.

Spontaneous Potential (SP) Log: Records natural electrical potential differences between formations and drilling fluid, offering insights into fluid content and formation boundaries.

Resistivity Logs: Provide data on the electrical resistivity of formations, helping to assess fluid saturation and identify hydrocarbon-bearing zones.

Density Log: Measures the bulk density of formations, aiding in porosity determination and lithology identification.

Neutron Porosity Log: Measures the hydrogen content of formations, assisting in porosity determination and fluid identification.

Sonic Log: Records the travel time of sound waves through formations, providing data for determining formation porosity, rock mechanical properties, and depth correlation.

Caliper Log: Measures the diameter of the wellbore, aiding in assessing hole conditions and well integrity.

These logs are essential for reservoir characterization, wellbore stability analysis, hydrocarbon exploration, and production decision-making.

figure 18: displaying well log file in vscode

The third 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 19: geologyc map of kensas city

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.

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).

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).

5- Verification and Downloading the Updated Data:

displaying the Data Frame (df) after filling missing values to visually confirm the changes.

Finally, the updated Data Frame is saved as a new CSV file named "updated_data.csv" using df.to_csv, and we can download it using Google Colab's download functionality (files.download).

6-Handling 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.

7-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 for inspection.

Finally, the normalized data is saved as a new CSV file named "normalized.csv" and offered for download.

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.