**KNIME ANALYTICS PLATFORM**



**TOLOSA, ANGELO P**

**DATASET**

This UN and IMF-sourced dataset was created as an easy-to-use and versatile source for many types of projects - social justice, economics, public policy, education, and more. When I posted my notebook about class inequality and university choice, I looked high and low for an economic classification dataset that I could merge with my base data, but couldn't find one - not even in the U.N. dataset collection. I made this to fill that need, and I hope it helps you out.

All data is sourced from the most recent info and numbers I could find from highly reputable sources. I'll definitely update it when new info comes available, so bookmark this and watch for updates.

Dataset Link:

<https://www.kaggle.com/datasets/jeanetteponder/world-economic-classifications>

**Dictionary**

The World Economic Classifications dataset is organized into nine columns:

|  |  |
| --- | --- |
| **Column** | **Description** |
| **country\_name** | Country name as commonly known. When applicable, any variations appear in parentheses after the most common name. |
| **un\_class\_2014** | Classifications from the Development Policy and Analysis Division (DPAD) of the Department of Economic and Social Affairs of the United Nations Secretariat (UN/DESA), 2014. It describes basic economic conditions and is based on World Economic Situation and Prospects (WESP) data. |
| **imf\_class\_2023** | Classifications from the International Monetary Fund (IMF)'s April 2023 report, "World Economic Outlook Database: Groups and Aggregates Information." |
| **g7** | Denotes countries that are part of the United Nations Gang of 7, or G7, and are ranked as Major World Economies by the IMF. |
| **eu\_member** | Member states of the European Union, both established and new. |
| **fuel\_exp\_country** | Fuel exporter country. According to DPAD, "An economy is classified as a fuel exporter if the share of fuel exports in its total merchandise exports is greater than 20 per cent and the level of fuel exports is at least 20 per cent higher than that of the country’s fuel imports. This criterion is drawn from the share of fuel exports in the total value of world merchandise trade." |
| **wealth\_rank** | Wealth ranking according to gross domestic product (GDP) at Purchasing Power Parity, or PPP. Worldometer states, "PPP takes into account the relative cost of living, rather than using only exchange rates, therefore providing a more accurate picture of the real differences in income." |
| **gdp\_ppp\_2022** | Per capita GDP at PPP as of 2022. |
| **gdp\_pc\_2022** | Per capita GDP as of 2022.` |

What is ***knime***?

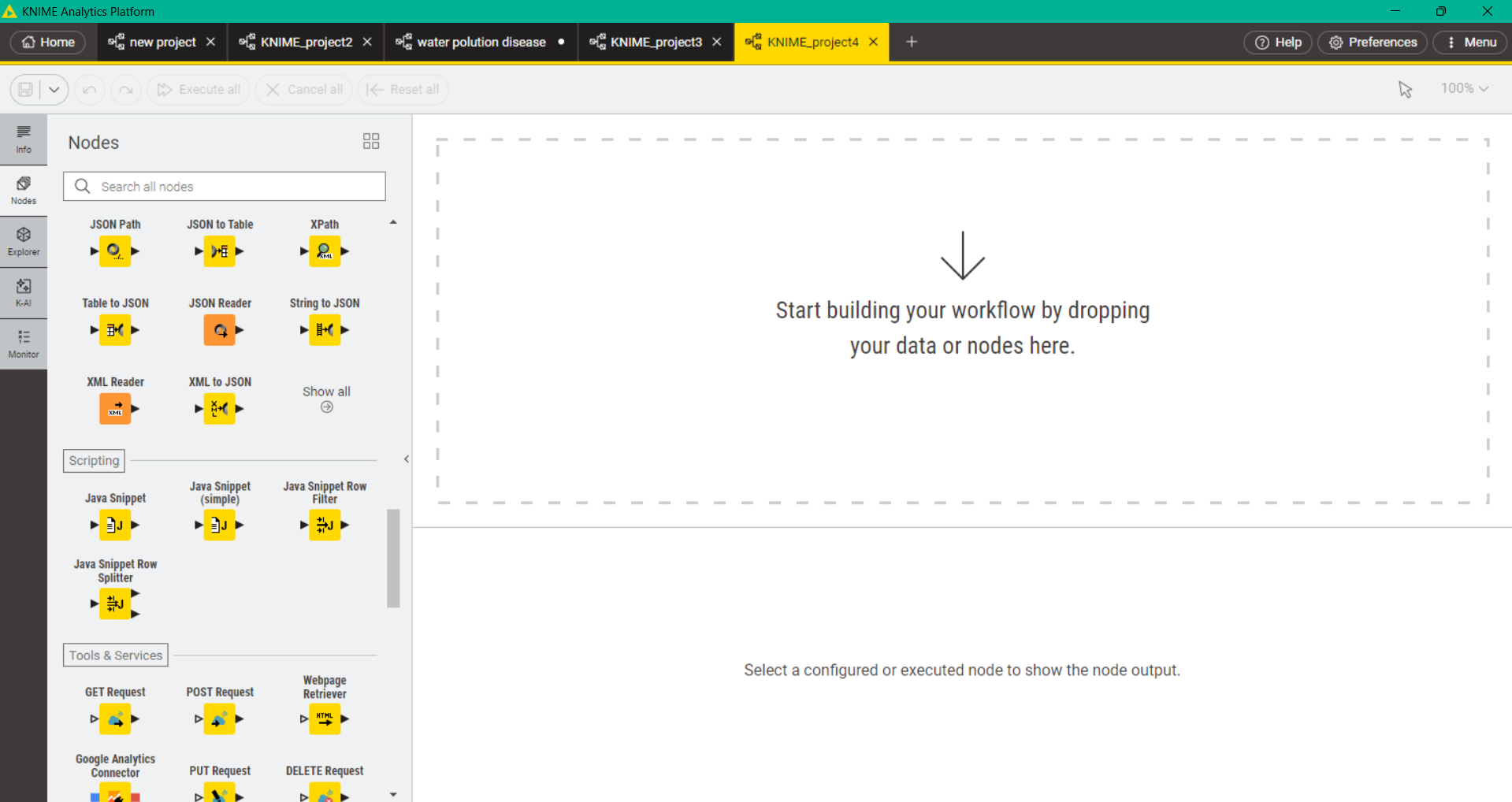
**KNIME** (Konstanz Information Miner) is an open-source data analytics platform that allows users to visually create data workflows by dragging and dropping pre-built nodes. It enables comprehensive data manipulation, analysis, and machine learning tasks with minimal coding, making it accessible for data integration, transformation, and predictive modeling across various industries.

**INSTALLATION:**

<https://www.knime.com/downloads>

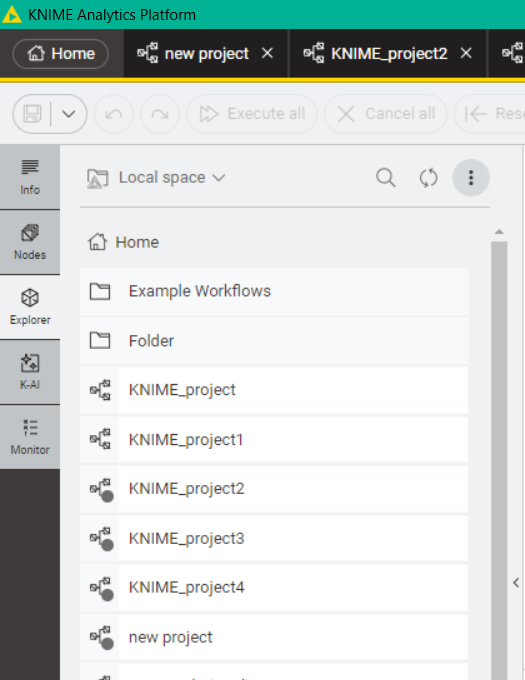
*(You need to sign in first before you download the app)*

After you install the app, it will lead you here:

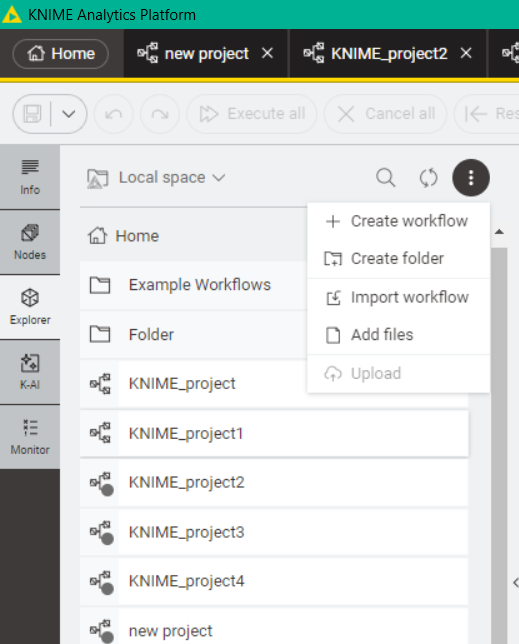


The logos on the left side are called nodes; each node has its own specific function. Some basic nodes include the 'File Reader', which is used to import data into your workflow; the 'Column Filter', which helps you remove or select specific columns; the 'Row Filter', which allows you to filter rows based on conditions; and the 'Data Viewer', which lets you preview the data. These nodes can be connected together to build a step-by-step data processing pipeline.

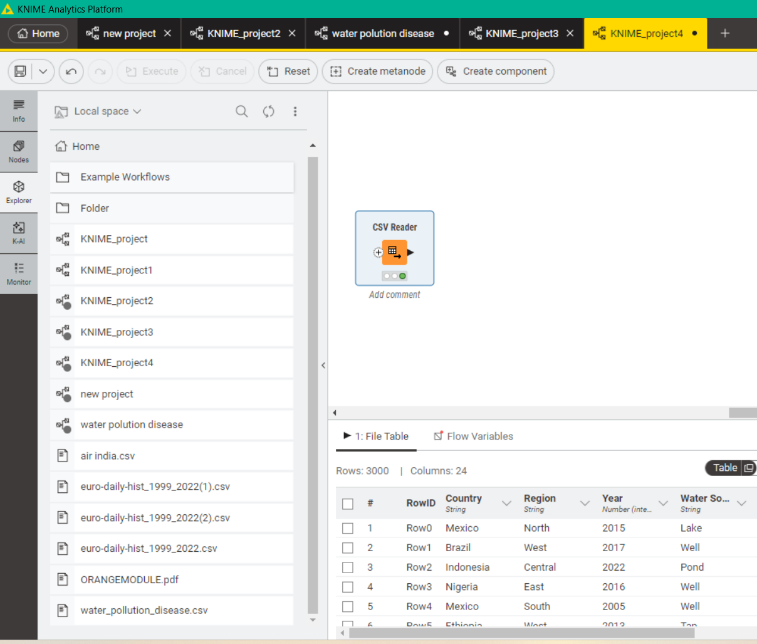
1. **CREATING NEW PROJECT**



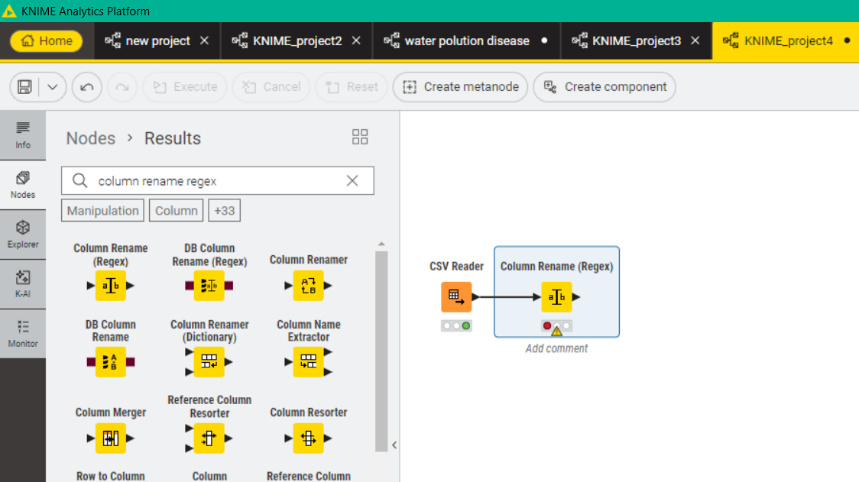
First, go to the left side of the screen and find the “Explorer” then the three dots beside the word Local space



Click the three dots and select add files. After you choose the data that you want to create it will appear under the Folder.



Drag your chosen data file to the open space and the reader nodes will appear as a representation of your file. After that you need to click the **play** button on the nodes to execute your file.



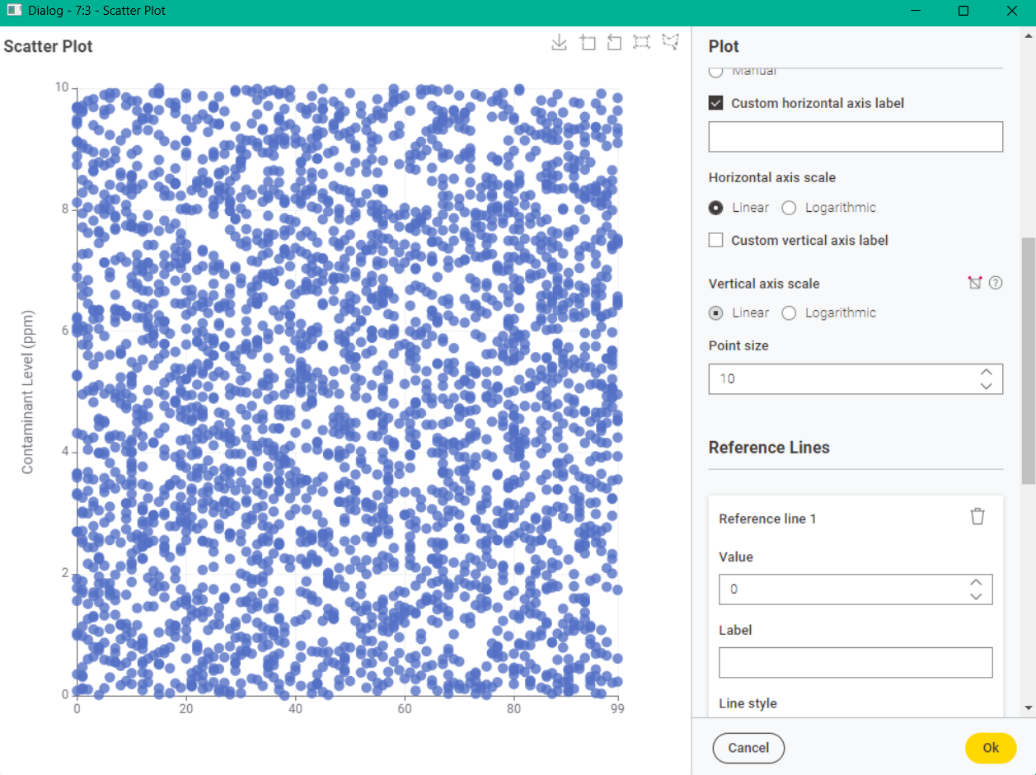
After you execute the file, you can now start to do your objectives to your data either for prediction, optimization, grouping/discovery, etc.

All you need to do is to find the node that suits to your objective, just drag the selected node that you want to execute next to the reader node. (For example, you want to clean your column names, you can add column regex to clear the excess expression in your column names) then you can proceed to another node that you need to complete your project.

(**NOTE**: THERE ARE SOME NODES THAT HAS MULTIPLE CONNECTOR THAT YOU CAN USE)

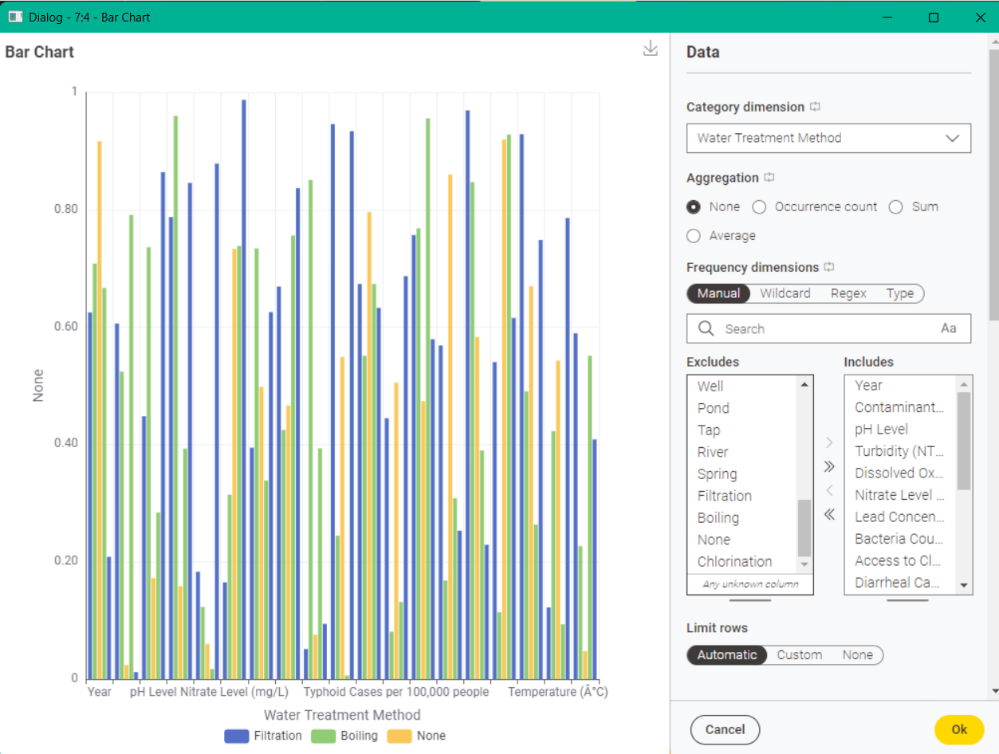
1. **DATA VISUALIZATION**

**SCATTER PLOT**



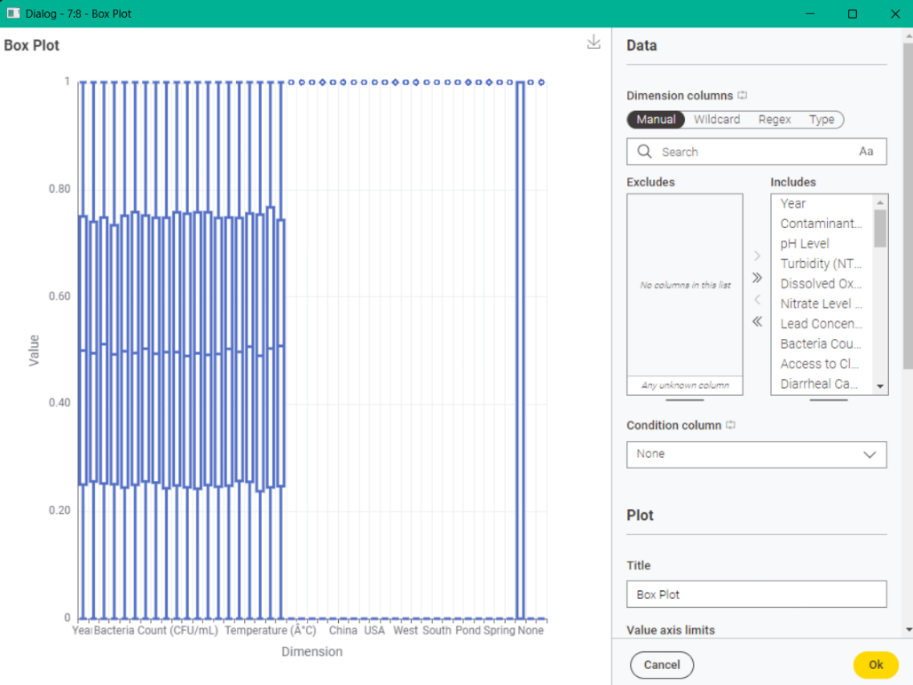
A **Scatter Plot** is a fundamental type of data visualization that displays the relationship between two numerical variables. Each point on the plot represents an observation (a row in your data), with its position determined by its values on the X and Y axes, it allows for immediate visual inspection of patterns, such as whether contaminant levels tend to increase with diarrheal cases. Furthermore, as a Plotly-based visualization, it offers powerful interactivity; users can zoom, pan, and hover over individual points to reveal detailed information from selected columns, enhancing data exploration. The configuration settings, visible in the image, allow for extensive customization, including specifying the numerical columns for the X and Y axes, mapping additional nominal or numerical columns to color, size, or shape of the points to visualize more dimensions, and customizing plot titles, axis labels, and tooltips for clarity and context.

**BAR CHART**



The **Bar Plot** node in KNIME is a versatile visualization tool used to compare values across different categories. Its primary help lies in effectively displaying the distribution, frequency, or aggregated measures of a numerical variable across distinct groups. This allows for quick insights into which categories are more prevalent, or which categories exhibit higher or lower average values for a particular metric. As a Plot-based visualization, it offers interactive features such as zooming, panning, and tooltips, enabling closer examination of individual bars and their corresponding data.

**BOX PLOT**



The **Box Plot** (*Plotly*) node in KNIME is primarily used to visually understand and compare the distribution of a numerical variable across different categorical groups. It effectively summarizes key statistical measures for each group, displaying the median, the spread of the middle 50% of the data (interquartile range or IQR), and identifying potential outliers. This allows you to quickly assess differences in central tendency, variability, and skewness of a numerical feature when broken down by distinct categories, For example comparing Contaminant Level distributions across various Water Source Types.

1. **DATA PREPROCESSING**

KNIME does not have a single, direct equivalent node that combines all those functionalities into one "reprocess" node.

**For Handling Missing Values:**

* **Missing Value node:** Used to replace, remove, or fill missing values in columns.

**For Scaling/Normalization:**

* **Normalizer node:** Used to scale numerical data (e.g., Min-Max normalization, Z-score standardization).

**For Encoding Categorical Features (One-Hot Encoding):**

* **One-to-Many node (or One Hot Encoder):** Used to convert nominal (string) columns into numerical (0/1) columns, which is required by most machine learning algorithms.

**For Discretization/Binning:**

* **Numeric Binner node:** Used to convert continuous numerical columns into categorical (binned) columns.

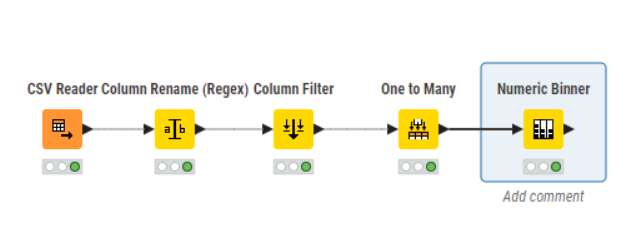
**For Type Conversion:**

* **String To Number node:** Converts string columns to numerical.
* **Number To String node:** Converts numerical columns to string.

**For Filtering/Selecting Columns:**

* **Column Filter node:** Selects or excludes specific columns from your dataset.

**FOR EXAMPLE:**



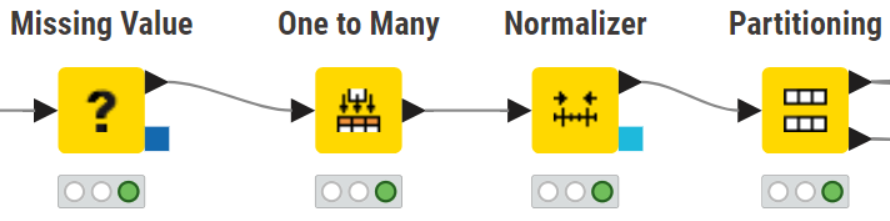
You need to be precise to every node that you’ll use because in other cases nodes cannot be executed because each node is designed for a highly specific function, meaning using the wrong one will prevent you from achieving your intended data transformation or analysis. This precision ensures data type compatibility between connected nodes, preventing common errors that arise from mismatched input/output expectations (e.g., expecting a string but receiving a number). Furthermore, correctly chosen nodes lead to accurate results, as they perform the exact operation required, unlike imprecise choices that might yield unintended or erroneous outputs. Finally, a precise workflow is inherently clearer and easier to understand, debug, and maintain, whether for your own future reference or for collaboration with others.

**PREDICTIVE ANALYTICS AND EVALUATION**

Predictive analytics is a type of data analytics that focuses on forecasting future outcomes or classifying new data based on historical patterns. In KNIME, this is achieved through the use of machine learning algorithms. You build predictive models using "Learner" nodes, such as the Random Forest Learner or Decision Tree Learner. These learned models are then applied to new data using "Predictor" nodes, like the Random Forest Predictor or Decision Tree Predictor, to generate the actual predictions. The performance of these models is then evaluated using "Scorer" nodes for classification tasks or "Numeric Scorer" nodes for regression tasks.

**“NOTE: THERE ARE ESSENTIAL NODES THAT YOU NEED WHEN PREDICTING USING KNIME”**

**1. Data Preprocessing & Manipulation**



* Missing Value: Essential for handling missing data.
* Column Filter / Row Filter: For selecting specific columns or rows based on criteria.
* Column Rename / Column Resorter: For organizing and naming your columns.
* Missing Value Column Filter: To remove columns entirely if they have too many missing values.
* Numeric Normalizer / Min-Max Normalization / Z-Score Normalization: For scaling numerical features, which is crucial for many algorithms.
* One to Many: To convert categorical features into numerical ones using one-hot encoding, necessary for most ML algorithms.
* Category to Number: Converts nominal (categorical) data into numerical representations.
* Math Formula / Column Expressions: For creating new features, applying mathematical operations, or complex conditional logic.
* Rule Engine: For creating new columns or filtering rows based on custom rules.
* GroupBy: For aggregating data based on group categories.
* Joiner / Concatenate: For combining data from multiple tables.

**2. Data Partitioning**



* Partitioning: This is arguably *the* most essential node for any supervised machine learning task. It splits your dataset into training and test sets (and optionally a validation set) to prevent overfitting and enable unbiased model evaluation. It offers options for random, stratified, and linear sampling.

**3. Machine Learning Algorithms (Learners & Predictors)**



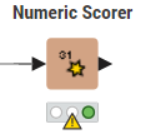
KNIME organizes most algorithms into a "Learner" node (for training) and a "Predictor" node (for applying the trained model).

* Classification:
* Decision Tree Learner / Predictor: A simple yet powerful algorithm, often used as a baseline.
* Random Forest Learner / Predictor: An ensemble method known for its robustness and accuracy.
* Gradient Boosting Machine Learner / Predictor (e.g., XGBoost, LightGBM, H2O GBM): Highly powerful ensemble methods, often top performers in competitions.
* Logistic Regression Learner / Predictor: For binary and multiclass classification, provides probabilistic outputs.

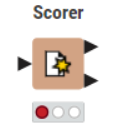
.

* SVM Learner / Predictor (Support Vector Machine): Good for high-dimensional data and clear margin separation.
* Naive Bayes Learner / Predictor: A probabilistic classifier, often a good baseline, especially for text classification.
* K-Nearest Neighbors (k-NN) Learner / Predictor: A simple, instance-based learning algorithm.
* Regression:
  + Regression Learner / Predictor: For modeling linear relationships.
  + Decision Tree Learner / Predictor (Regression): Decision trees can also be used for regression tasks.
  + Random Forest Learner / Predictor (Regression): Random Forests are very effective for regression.
  + Gradient Boosting Machine Learner / Predictor (Regression): Powerful for complex regression tasks.
  + Numeric Scorer (for regression evaluation): To calculate metrics like R², MSE, MAE, RMSE.
  + Clustering (Unsupervised Learning):
  + K-Means: A popular algorithm for partitioning data into K clusters.
  + Hierarchical Clustering: Builds a hierarchy of clusters.

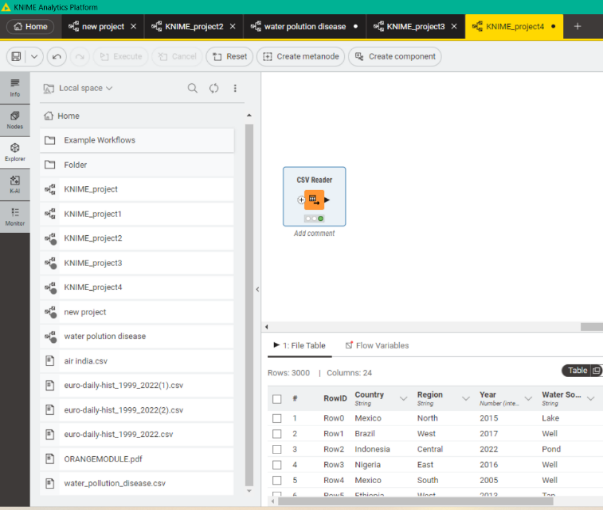
4. Model Evaluation

****

* Scorer: For classification models, calculates accuracy, precision, recall, F1-score, Cohen's Kappa, and generates confusion matrices.
* Numeric Scorer: As discussed, for regression models, calculates R², MSE, MAE, RMSE, and MAPE (with the caveats discussed).
* ROC Curve (ROC Curve): For evaluating binary classifiers, plots the True Positive Rate vs. False Positive Rate.
* Lift Chart: Visualizes model performance, especially in marketing/CRM applications.

****

**CREATING MODEL**

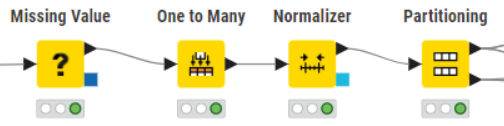
**STEP 1:**

* **Import your selected data using reader nodes**

**s**

**Step 2:**

* **Manipulating data (Initial data cleaning): this nodes helps you to determine your dataset error like missing values, column name error, etc. There are many manipulation nodes that you can use that satisfy the needs of your dataset.**

****

* Missing Value
* Missing data can cause errors or inaccurate results in subsequent analytical steps and machine learning models, so this node ensures data completeness and integrity.

One to Many

* Most machine learning algorithms (especially those based on mathematical operations like regression or neural networks) cannot directly process text-based categorical data, requiring numerical input.

Normalizer

* Many machine learning algorithms are sensitive to the scale of input features; features with larger ranges can disproportionately influence the model. Normalization ensures all features contribute fairly and can speed up model convergence.

**(NOTE: THIS MANIPULATION NODES IS SUITABLE IN MY MODEL THAT’S WHY THEY HERE BUT IN SOME INSTANCES, YOU NEED TO ADD OR REMOVE SOME NODES TO SATISFY THE NEEDS OF YOUR DATASET)**

**Step 3:**

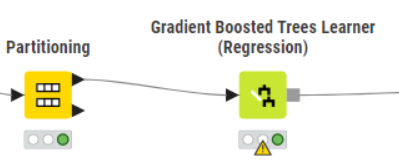
* **ADD PARTITIONING NODE FOR SPLITTING RATIO (the standard is 80% and 20% for training and testing respectively.)**



* Partitioning node allows you to configure the split ratio (e.g., 80% for training, 20% for testing) and offers options like "Stratified sampling". Stratified sampling is particularly useful for classification tasks, as it ensures that the proportion of your target classes is maintained similarly in both the training and test sets, which is important for accurate evaluation.

**Step 4:**

* **ADD GRADIENT BOOSTED TREES LEARNER (REGRESSION) NODE TO TRAIN YOUR DATA AND SELECT YOUR TARGET COLUMN**

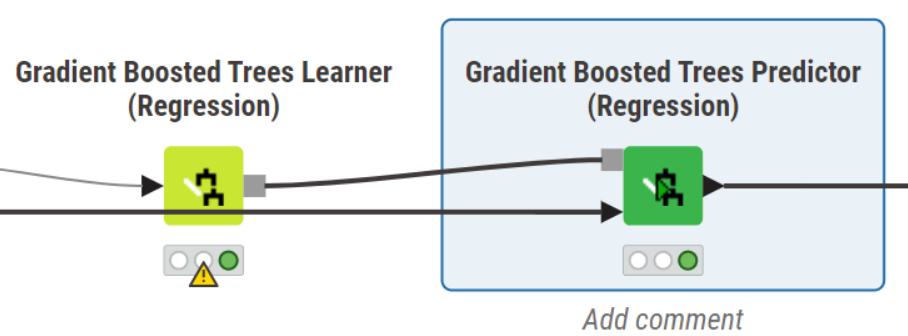
****

Training your data using a Gradient boosted Regression (Regression) is important when your goal is to predict a continuous numerical value (e.g., predicting exact contaminant levels, specific GDP figures, or rainfall amounts) rather than a category.

**(NOTE: THERE ARE TWO DIFFERENT LEARNER NODE FOR EVERY ALGORITHM THE LEARNER FOR CLASSIFICATION AND THE LEARNER FOR REGRESSION**

**Step 5:**

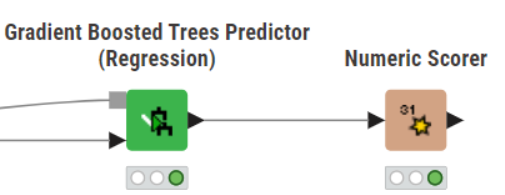
* **ADD GRADIENT BOOSTED PREDICTOR (REGRESSION) NODE TO SELECT PREDICTION COLUMN (The target column for prediction that I choose is  “GDP\_PC\_2022”.**

****

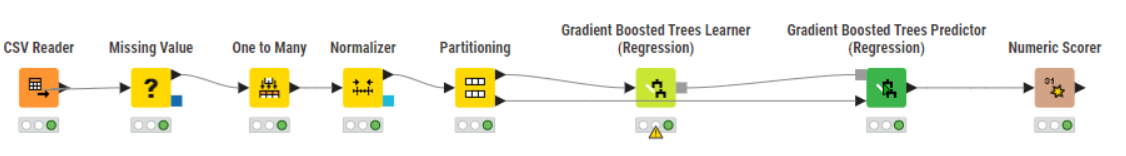
The Gradient Boosted Trees Predictor (Regression) node in KNIME is used to apply a previously trained Gradient Boosted Trees regression model to new, unseen data in order to make continuous numerical predictions. It takes the "knowledge" learned by the Gradient Boosted Trees Learner (Regression) node and uses it to estimate a numerical target value for each row in the input data.

(NOTE: YOU NEED TO CONNECT BOTH PARTITION AND GRADIENT BOOSTED LEARNER (REGRESSION) NODE ON THE GRADIENT BOOSTED TREE PREDICTOR BECAUSE THEY SERVE TWO DISTINCT AND ESSENTIAL PURPOSES FOR MAKING PREDICTIONS).

**Step 6:**

* **LASTLY, ADD A NUMERIC SCORER TO CALCULATE STANDARD SET OF METRICS SPECIFICALLY DESIGNED FOR REGRESSION TASK.**

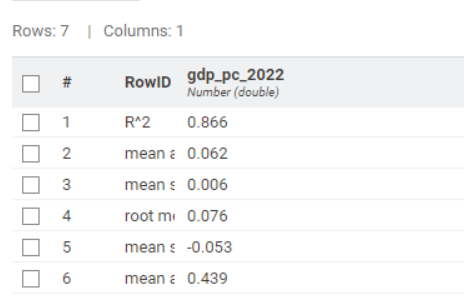
The Numeric Scorer node is fundamentally important in a regression workflow because it provides the objective, quantifiable evidence of your model's performance. Without it, you would simply have a list of predictions without knowing how accurate or useful they truly are.

**RESULT**

After you do all the step your workflow will look like this. All you need to do is to execute the numeric scorer to view the value of the metrics under your workflow.

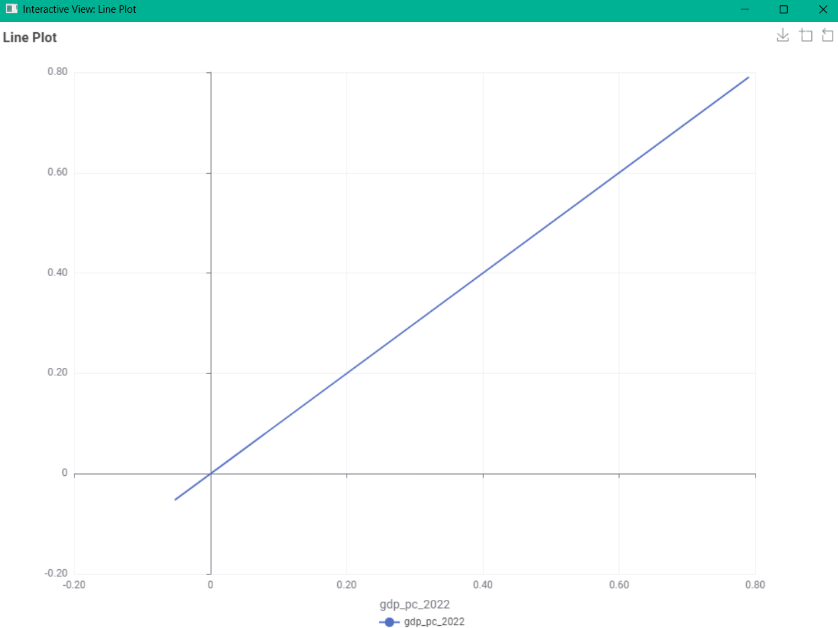
**This is the final result of the model:**

**We calculated some standard metrics needs for prediction (R^2, MAPE, MAE, MSE, RMSE, MSD, ADJUSTED R^2).**

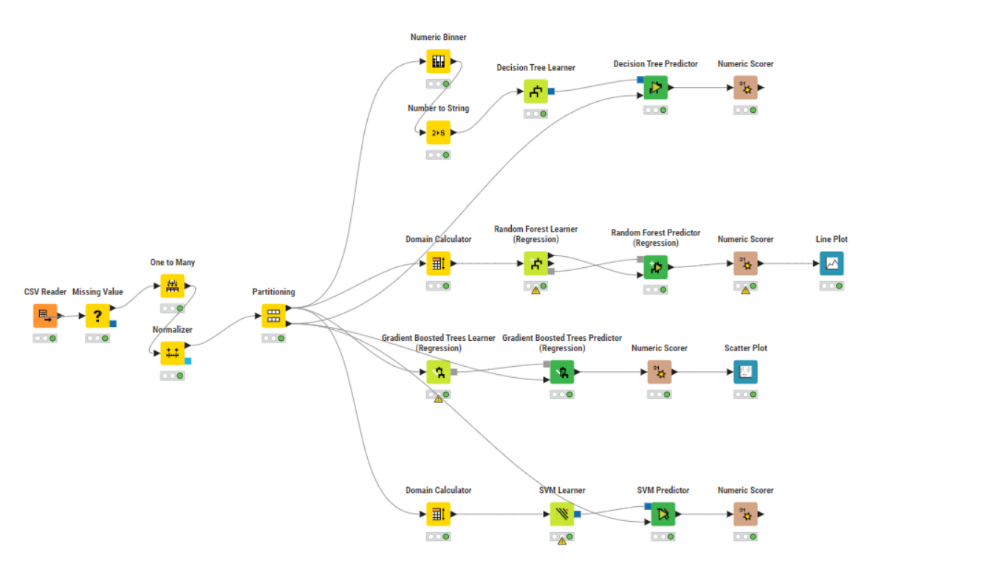
****

**GRAPHING THE RESULT:**

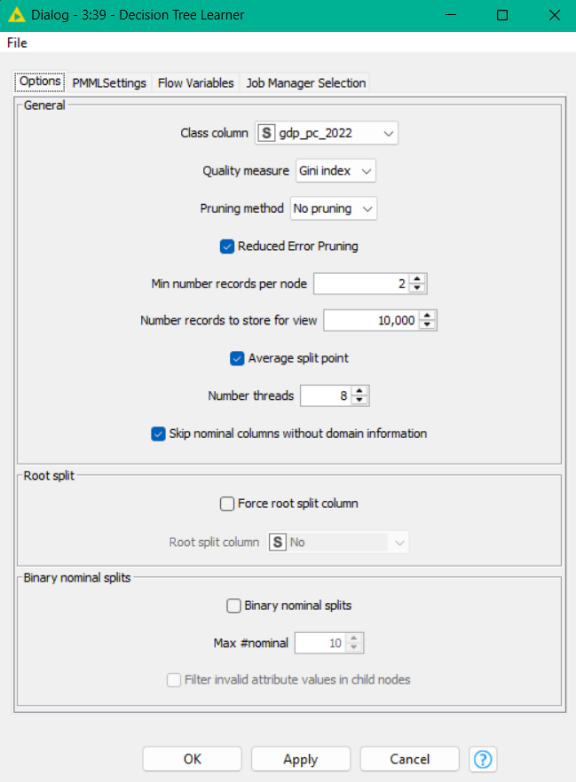
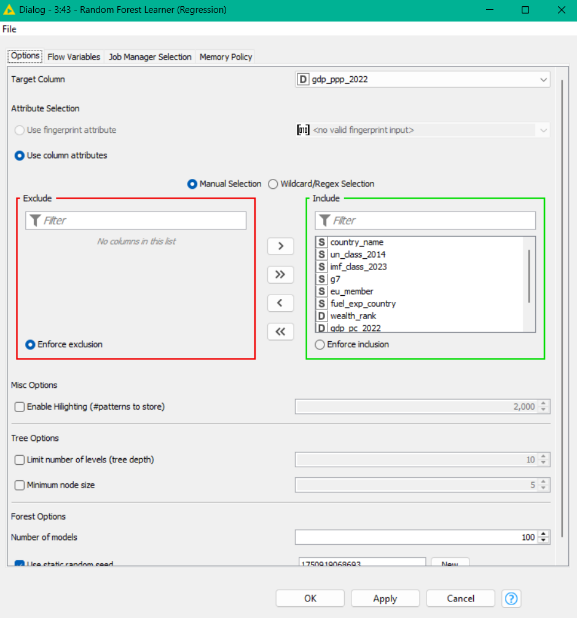
GO TO THE “VIEWS” NODES TO SELECT WHICH VISUALIZATION DIAGRAM YOU WANT TO USE.

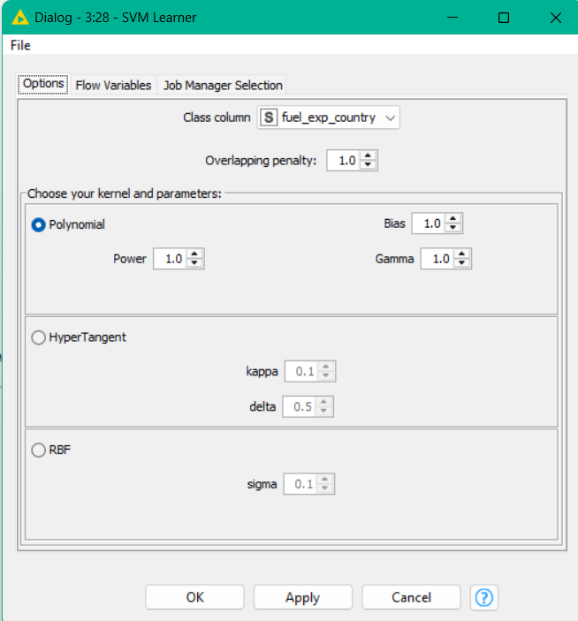
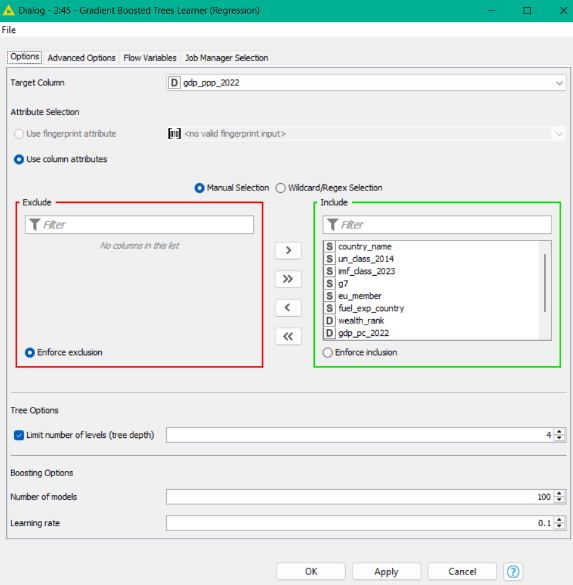


**WE USE DIFFERENT KIND OF MODELS FOR BETTER COMPARISON TO EACH RESULT**



**THESE ARE THE SOME SETTINGS THAT I USED EVERY LEARNERS NODES FOR THE MODELS…**





**(NOTE: IN KNIME, THEY HAVE THE SUGGESTED SETTINGS YOU CAN USE THE LEARNERS NODE WITHOUT MANIPULATING THE DEFAULT SETTING OF THE LEARNERS NODES)**

**ADVANTAGES AND DISADVANTAGES OF USING KNIME:**

**Advantages of using KNIME:**

* **Visual Workflow and Low-Code/No-Code:** Offers an intuitive drag-and-drop interface, making it accessible for building data pipelines without extensive coding.
* **Modular Design and Granular Control:** Each node has a specific function, allowing for precise control over every step of data processing and analysis.
* **Comprehensive Functionality:** Provides a wide array of tools for data access, transformation, feature engineering, machine learning (both learning and prediction), and interactive visualizations.
* **Strong for Data Preparation/ETL:** Excels in cleaning, structuring, and transforming raw data for analysis.
* **Interactive Visualizations:** Offers dynamic and interactive plots (e.g., Plotly-based) that enhance data exploration.
* **Transparency and Debugging:** The visual nature of the workflows makes it easier to understand data flow and identify/troubleshoot errors.

**Disadvantages of using KNIME:**

* **Learning Curve for Node Configuration:** While visual, configuring individual nodes precisely for specific tasks can require some learning and understanding of their nuances.
* **Multi-Node Solutions for Integrated Tasks:** Combining diverse analytical results or performing highly integrated operations often requires chaining multiple specific nodes, rather than a single, all-encompassing node.
* **Iterative Troubleshooting:** Diagnosing and resolving complex issues, particularly those related to data types or upstream node configurations, can sometimes be an iterative process.
* **Explicit Data Type Management:** Users often need to manually manage and convert data types between nodes to ensure compatibility for downstream operations.