Optimizing the Output of a Remote Sensing Physical Model using Machine Learning Algorithms

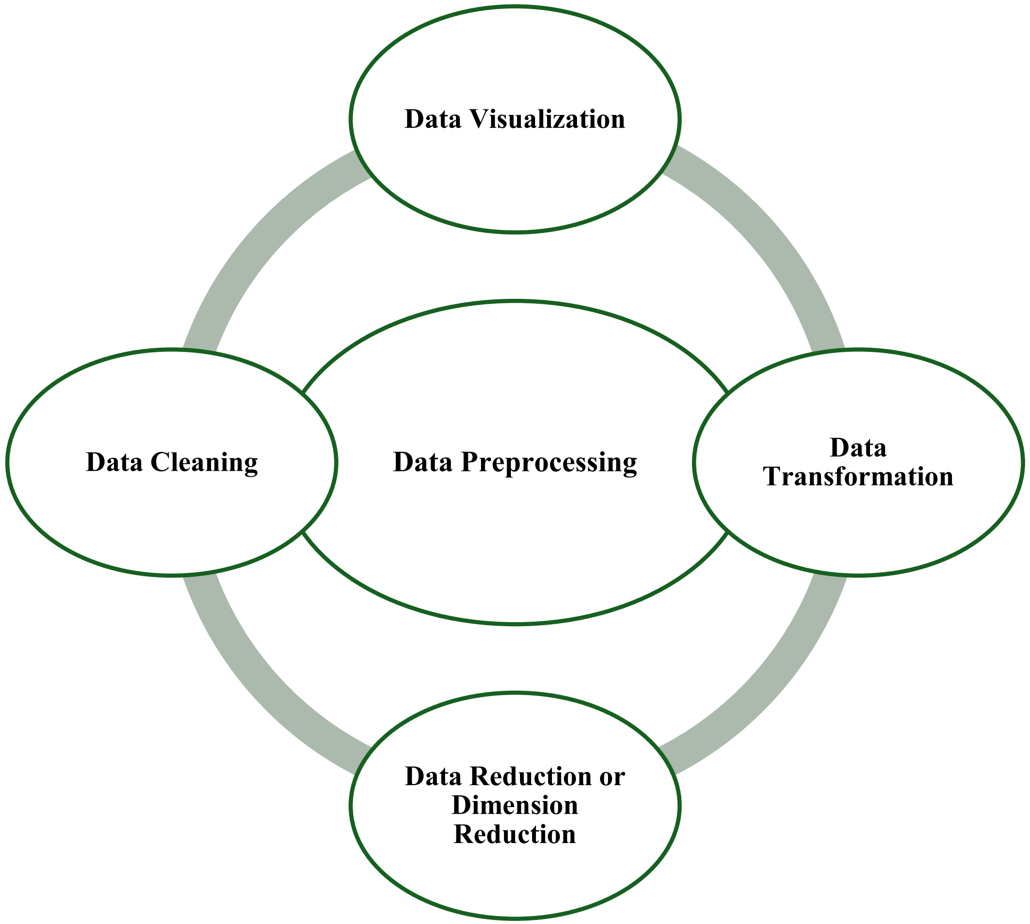
Alvin Tung Wei Lik
Mentor: Li Fang

Objectives

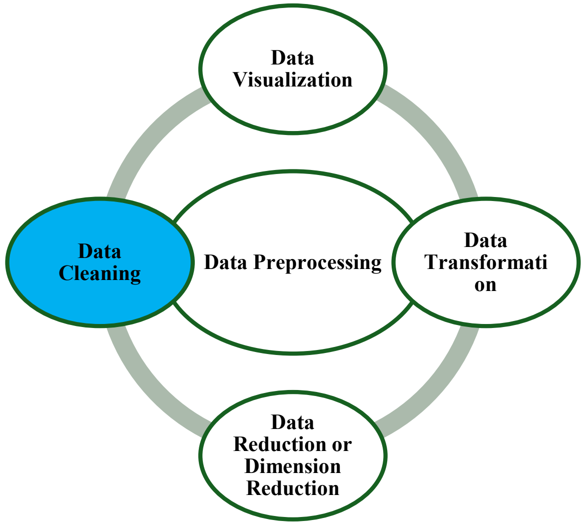
- The objective of the project is to optimize the output of a remote sensing physical model using machine learning algorithms, resulting in smoother predictions with lower noise

- This will be achieved through data preprocessing, fine-tuning model parameters, selecting appropriate algorithms, and incorporating noise-reduction strategies to ensure that the model delivers more consistent and dependable results.

Data Flow

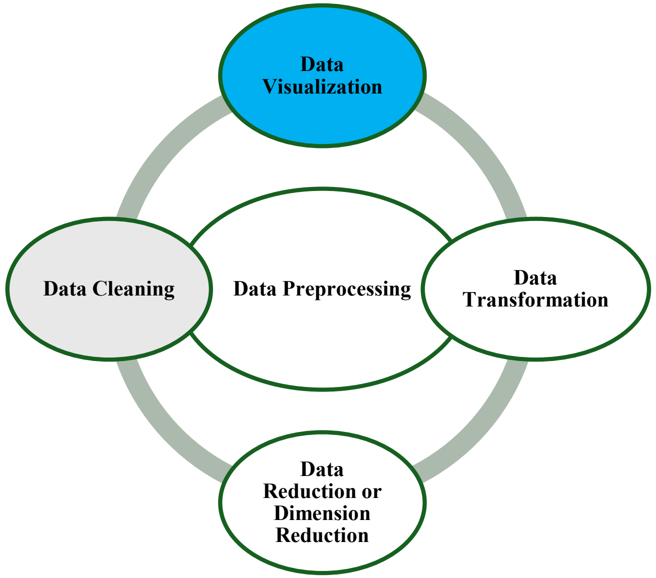

Data Visualization
Data Cleaning
Data Preprocessing
Data Transformation
Data Reduction or Dimension Reduction

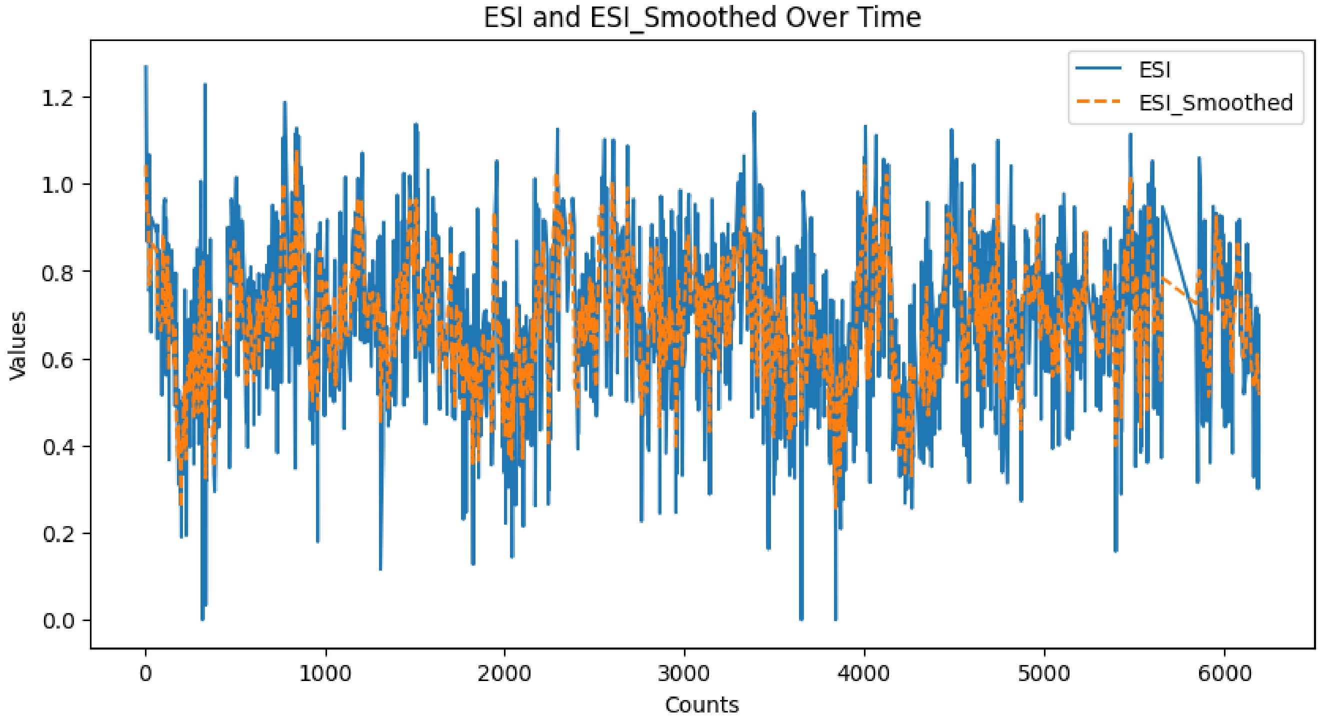
Data Cleaning



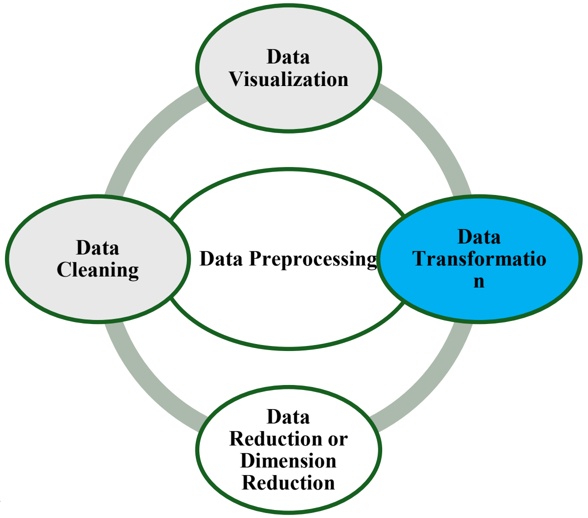
- Analyzed data on the datasets provided by my mentor (Li Fang)
- Remove the null values so that the dataset is ready to use

Data Visualization



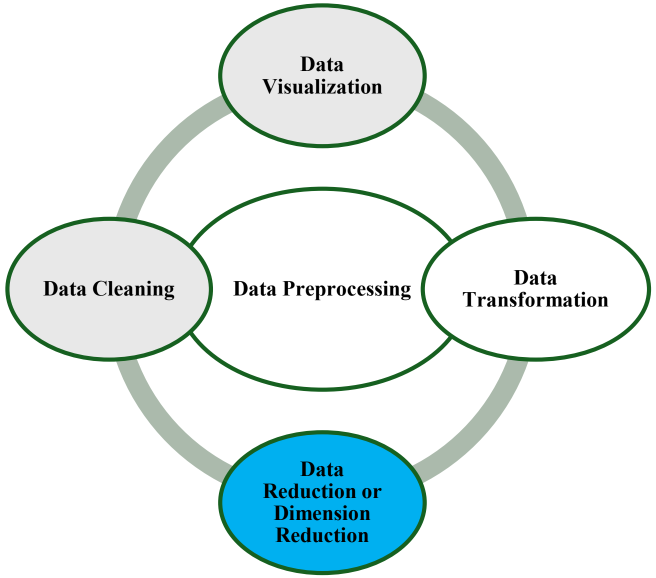


Data Transformation



- Data Cleaning: Removing null values to ensure the dataset is clean and ready to be use
- Data Normalization: Adjusting data to a common scale or format to ensure consistency
- Data Merging: Combining data from different sources or datasets into a single dataset
- Data Filtering: Selecting a subset of data based on specific criteria

Data Reduction or Dimension Reduction



- Reduce the number of columns to reduce the complexity of the model
- Less redundant data means the model is likely to learn from noise
- The model becomes easier to interpret and may generalize better

Result for different algorithm

```
Score of Linear Regression
Mean Squared Error : 0.008444429092454601
R^2 Score: 0.5040673331053765
```

```
Score of Decision Tree Regression
Mean Squared Error : 0.012524965447056004
R^2 Score: 0.2644216146628875
```

```
Score of Random Forest Regression
Mean Squared Error : 0.012524965447056004
R^2 Score: 0.2644216146628875
```

```
Score of XGBoost Regression
Mean Squared Error : 0.009525491891077579
R^2 Score: 0.4405776227967527
```

Conclusion

- **Real-World Data:** Significantly larger and more complex than school-provided datasets.
- **Exploratory Data Analysis (EDA):** It is crucial before building a model.
- **Evaluation:** Check training and testing scores.
- **Metrics Analysis:** Use confusion matrix, accuracy score, etc., to assess model performance.
- **Tuning:** Adjust the model to improve performance.
- **Overfitting and Imbalance:** Strategies to control these issues.