**Step-by-step process for data engineering and modelling**

1. Project goal

To Complete the WMO's Heatwave Warning Systems Inventory by filling in missing data (yes, no, unknown) using data engineering and machine learning techniques.

2. Data Collection with Python and BeautifulSoup:

- Web Scraping: Utilized Python and BeautifulSoup to scrape data from Google Scholar and perform Google searches for heatwave warning systems.

- Extracted Information: Used BeautifulSoup to extract relevant information from web pages, including keywords and countries mentioned.

- Filtered URLs: Employed Python to filter out web pages where country data couldn't be obtained.

3. Data Preprocessing:

- Matching Data: Matched web-scraped data with WMO's dataset to categorize entries as known or unknown.

- Keyword Analysis: Employed Python to create binary labels (0 for not present, 1 for present) based on keyword presence.

- Class Imbalance: Addressed class imbalance issues using Python libraries such as SMOTE.

4. Model Selection:

- Python Libraries: Utilized Python libraries -Scikit-Learn to explore various classification machine learning models

- Ranking Models: Determined model selection based on accuracy scores.

- Selected Random Forest Classifier as the final model.

5. Model Training and Testing:

- Python for Model Training: Trained the Random Forest Classifier using known data with Python.

- Cross-Validation: Employed Python libraries to perform cross-validation and hyperparameter tuning.

- Testing with Python: Applied the trained model to predict heatwave warning system status for unknown entries and evaluated model performance using Python tools.

- Accuracy of Random Forest Classifier: 0.7042

6. Model Evaluation (Classification Report):

- Precision and Recall: The precision for class 0 is 0.67, class 1 is 0.72. Recall for class 0 is 0.53, class 1 is 0.82.

- F1-Score: The F1-score for class 0 is 0.60, class 1 is 0.77.

- Support: The support for class 0 is 58 and for class 1 is 84.

- Overall accuracy: 0.70

- Macro-averaged F1-score: 0.68

- Weighted-averaged F1-score: 0.70

7. Results:

Achieved an accuracy of 0.7042 in predicting heatwave warning systems, with a balanced combination of precision, recall, and F1-scores.

8. Challenges:

- Class Imbalance: Addressed class imbalance challenges using Python's SMOTE technique.

- Web Scraping Complexities: Overcame web scraping complexities with BeautifulSoup to obtain relevant data and handled noisy data using Python.

lts.