Week 3: Unsupervised Learning and Feature Engineering

Task 3.2: Feature Engineering

Objective: Develop skills in creating and selecting features to improve machine learning model performance.

Dataset: *Synthetic dataset using sklearn's* `make_classification`. This allows you to generate a large random dataset tailored to the task, with control over the number of informative, redundant, and noisy features.

- **Dataset Generation Command**: You can create this dataset directly in your Python environment using the following sklearn command:

```
from sklearn.datasets import make_classification

X, y = make_classification(n_samples=1000, n_features=20, n_informative=2, n_redundant=10, n_clusters_per_class=1, weights=[0.99], flip_y=0, random_state=1)
```

Activities:

1. **Feature Creation**:
    - Generate the dataset and add manually created features that could be interaction terms or polynomial features.
2. **Feature Selection**:
    - Implement various feature selection techniques to determine the most impactful features.
    - Evaluate the effect of feature selection on model performance.
3. **Model Building**:
    - Rebuild classification models using selected features to compare performance against the baseline model with all features.

Expected Output:

- A Jupyter notebook demonstrating the feature engineering process, including the impact of selected features on model performance.

Documentation:

- Detail the feature engineering process, selection methods used, and effects on the predictive model.