**Week 2 Task Sheet: Feature Engineering & Initial Model Development**

**Objective:**

By the end of this week, students should have engineered relevant features from climate and yield data, selected target variables, and implemented baseline deep learning models to start predicting crop yields. This sets the stage for iterative model improvements in subsequent weeks.

**Tasks to Complete This Week:**

Download HRRR\_01\_AL\_2022 ( Alabama State.) dataset for all 12 months of year 2022 and USDA\_Corn\_County\_2022.csv.

Make sure you’ve:

* Merged both datasets by state and county
* Parsed the date columns into a proper datetime format
* Handled missing or inconsistent values

1. **Feature Engineering**

**Climate-Based Features (From HRRR Data)**

**Aggregate Daily Weather Features**

Group by county and date to calculate:

* Daily max temperature
* Daily min temperature
* Daily average temperature
* Daily total precipitation

**Seasonal or Growing Period Aggregates**

Sum or average over the growing season (e.g., April–September):

* Avg Growing Season Temp (mean daily average temp)
* Cumulative Rainfall
* Number of Hot Days (e.g., temp > 35°C)
* Number of Cold Days (e.g., temp < 10°C)
* Days Without Rain (precipitation == 0)

**Derived Indicators**

**Growing Degree Days (GDD):**

GDD=max(0,Avg Daily Temp−Base Temp)

Common base temp is 10°C for most crops.

**Rainfall Variability:**

* Standard deviation of daily rainfall across the growing season.

**Heat Stress Index:**

* Number of consecutive days above a certain threshold (e.g., 33°C).

**Drought Periods:**

* Maximum length of consecutive days with little or no rain (e.g., <1mm).

**Crop-Based Features (From USDA Yield Data)**

From USDA\_Corn\_County\_2022.csv:

* Crop Yield (Target): Likely already present (e.g., bushels/acre)
* Normalized Yield: z-score or percentile within state or region
* Region Encoding: Convert county and state to categorical labels or embeddings
* Soil Type Proxy: If soil data not available, use historical average yield as a proxy
* Previous Year Yield: Lag feature if time-series data is available

**Final Feature Matrix**

Your final dataset should have:

* 1 row per county per year (or season)
* Climate features: avg temp, rainfall, GDD, heat days, drought periods, etc.
* Crop features: yield, normalized yield, prior yield (if available)
* Optional: satellite NDVI or vegetation indices (if added later)

**Preprocessing**

* Normalize all numerical features (z-score or min-max scaling)
* Encode categorical features (state, crop type) using:
* Label Encoding or
* One-Hot Encoding for small sets
* Check multicollinearity to remove redundant features before modeling

1. **Define Prediction Targets**
   * Define what you’re predicting: e.g., crop yield (bushels/acre) or yield deviation from the mean.
   * Justify the choice of prediction target.
   * Ensure that the prediction target is clearly separated from input features to avoid data leakage.
2. **Split the Data**
   * Split the dataset into training and test sets (e.g., 80/20).
   * Consider also preparing a validation set (10–20%) for hyperparameter tuning.
3. **Build a Baseline Model**
   * Implement a simple feedforward neural network using TensorFlow or PyTorch.
   * Use a small architecture (1–2 hidden layers, ReLU activation).
   * Train the model and track metrics such as MSE, MAE, and R².
4. **Evaluate Model Performance**
   * Compare actual vs. predicted yield using line or scatter plots.
   * Report key performance metrics on training and test sets.
   * Reflect on whether the baseline model captures any meaningful patterns.
5. **Prepare for Next Week**
   * List 2–3 ideas for improving the model (e.g., better features, different architectures).
   * Think about how weather extremes might be modeled better using time-series data or CNNs for spatial analysis.

**Deliverables by End of Week 2:**

* Cleaned and finalized dataset with engineered features.
* List and description of all features used.
* A baseline deep learning model implemented and evaluated.
* Performance report with graphs and error metrics.
* 1–2 paragraph reflection on lessons learned and modeling ideas for Week 3.