**Clustering**

**Tried Tslearn DTW method and Kmeans clustering** (form sklearn with euclidean distance) **for height LA and biomass (both log-transformed and raw data, and also tried on simulated data)-> does not find meaningful clusters**

Simulated data is created by randomly picking three genotypes to take all the plants with those genotypes -> duplicate three times (or also add random noise or shift on days)

**Prediction with DH in the platform**

1. [Identify missing value mechanism](23_02meeting.pptx)s. **The little test of missing value is not MACR**

**2. Take the predicted value, derive and derive2 from p spline of LA and Height (because an error occurred when I applied it on biomass) as input X; Randomly pick SNPs every 1cM (1116 SNPs were chosen)**

**Log transformed data**

**Remove single data point outlier**

**Remove time series outliers for genotype with 3 rep**

**Spatial correction**

**Fit spline (knot larger-> more smooth)**

**3. only take the plants with LA and Height information after removing the outlier**

**4. scaling-**> default setting

Use StandardScaler from sklearn (Standardize features by removing the mean and scaling to unit variance. z = (x - u) / s)

**5.two layer lstm + a linear layer + sigmoid:**

Input: torch.Size([880, 43, 6])  
LSTM\_classification(

(lstm): LSTM(6, 2, num\_layers=2, dropout=0.01)

(fc): Linear(in\_features=2, out\_features=1, bias=True)

(sigmoid): Sigmoid()

)

**6. find SNPs, whose model at least has three out of four TP, FP, TN, and FN.** (which means the model is not just simply assign all the plants with the same label) -> 37 SNPs out of 1116 are found -> notice there are some NA in the confusion matrix (imbalanced in SNPs labels, which may cause all data points with label 1 to be in train but not in test) -> need prefiltering instead of randomly chose SNPs based on distance

**Use biomass simulated data**

Use SDE to generate 100 time seriers with different r, use the simulated data as input X and use r as input Y

Use 80:20 training test split to test if lstm model can predict the r based on input time seriers data. The noise we use her is 0.05 \* sqrt(t), t is the time step(day at this case) and we use the original biomass without devided by 1000