**Q2 Report**

In this question, we are required to determine whether the external factors are affecting the yield of the FFB or not. The variables ‘HA Harvested’ and ‘FFB\_Yield’ are provided in the data. However, the ‘FFB\_Yield’ provided is not the actual yield, but is normalised by the area harvested. Therefore, I excluded the ‘HA\_Harvested’ from subsequent analysis as it is just the base number of ‘FFB\_Yield’ (higher area most probably would end up with lower yield per area).

We can first study the correlation of the different variables with the FFB\_Yield using the time series plot, as shown below:

A picture containing chart

Description automatically generated

The variables are normalised to [0, 1] range so that they can be visualised easily in a single plot. From here, we can observe whether any of the variables had similar fluctuation as the ‘FFB\_Yield’. The ‘FFB\_Yield’ actually exhibited noticeable seasonality, which somehow matches well with ‘Precipitation’ and ‘SoilMositure’.

Then, we also can plot a correlation matrix to check the correlations between each pair of variables. Three types of correlation are used: Pearson, Spearman and Kendall.

Chart, waterfall chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Here, we can see that the ‘FFB\_Yield’ had the highest correlation with the ‘Precipitation’, despite it is apparently weak. Therefore, I proceed to use the neural network (a simple MLP) and linear regressor (non-linear vs. linear), coupled with SHAP Analysis to confirm this relationship. The MLP and linear regressor have training error of approximately 10 – 15 % (random state not set to ensure reproducibility).

By using the SHAP analysis, I obtain the following figures:

Chart

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

For both NN and linear models, we can observe that the ‘Precipitation’ and the ‘SoilMoisture’ are the two most influential variables to predict or estimate the ‘FFB\_Yield’. In fact, both agree that higher ‘Precipitation’ and lower ‘SoilMoisture’ are favourable for higher ‘FFB\_Yield’. Somehow the influence of the ‘SoilMositure’ defies our perception. Cited from a recent research study, ‘local water stress is then enhanced, favoring a reduction in total yields’. Thus, the variables provided by the dataset alone potentially are insufficient to make conclusive remarks. More water stress related metrics, such as the evapotranspiration and drought indices shall be included.

Nevertheless, from the linear model, we find that the importance of the variables can be arranged according to the coefficient, as shown in the table below.

|  |  |
| --- | --- |
| **Variable** | **Coefficient** |
| ‘Precipitation’ | 0.6277 |
| ‘SoilMoisture’ | -0.2566 |
| ‘Working\_days’ | 0.0587 |
| ‘Min\_Temp’ | -0.0565 |
| ‘Max\_Temp’ | 0.0065 |
| ‘Average\_Temp’ | 0.0018 |

References:

Oettli, P., Behera, S.K. and Yamagata, T., 2018. Climate based predictability of oil palm tree yield in Malaysia. *Scientific Reports*, 8(1), pp.1-13.