A. Which feature set combination worked best and which was the worst?

Best Feature Set Combination:

The best-performing feature set combination was [1, 4], with an accuracy of 0.8921 and a TSS score of 0.7805. This indicates that combining the core feature set (FS-I) with the max-min feature set (FS-IV) provides the best predictive performance. FS-I includes the main solar activity features like the magnetic field, while FS-IV captures the difference between maximum and minimum values of key parameters within a specific time window. This combination likely works well because it combines both core solar features and their dynamic variations, which are critical for predicting solar flares.

Worst Feature Set Combination:

The worst-performing feature set combination was [3], with an accuracy of 0.4413 and a TSS score of 0.0000. This suggests that the model was unable to make any meaningful predictions with just FS-III (the historical activity feature). This is because FS-III alone may not provide sufficient real-time information for the prediction task, as it only captures historical activity without incorporating key magnetic or temporal changes happening prior to the solar flare events.

Dataset Explanation:

Upon investigation, it appears that Dataset 3 (FS-III) effectively had no meaningful values. The absence of significant data likely caused the classifier to perform no better than random guessing, resulting in a TSS score of zero. This further reinforces the idea that FS-III alone is insufficient for accurate solar flare prediction, and it is best used in conjunction with other features (as seen in combinations like [1, 2, 4], where FS-III is included but combined with other more robust feature sets).

b. Does adding additional FS-III and FS-IV features improve the TSS score?

When considering whether adding additional features, particularly FS-III (Historical Activity) and FS-IV (Max-Min Feature), improves the True Skill Statistic (TSS) score, the results are somewhat mixed.

Impact of FS-III:

FS-III represents the historical activity of solar regions, capturing past solar flare activity in each region. However, the data from FS-III on its own does not seem to contribute any significant value to the model. For example, when FS-III is used alone (Combination [3]), both the accuracy (0.4413) and TSS (0.0000) plummet, making it clear that this feature set has little to no predictive power on its own. This is because FS-III was essentially empty or lacked meaningful data, resulting in the classifier's failure to make any informed predictions. The absence of valuable data from FS-III means that it does not bring any additional predictive information to the table, and using it alone leads to poor results.

Impact of FS-IV:

FS-IV, on the other hand, captures the differences between maximum and minimum values of several key solar parameters over time (e.g., magnetic field strength, current helicity). These time-based fluctuations are useful for detecting solar events since they

indicate the changing dynamics of solar regions that might lead to flares. Combinations that include FS-IV tend to perform better. For example:

- Combination [1, 4] (FS-I + FS-IV) yields a TSS score of 0.7805, the highest among all combinations.
- Even combinations like [1, 2, 4] and [1, 3, 4] (which include FS-IV) show relatively high TSS scores of 0.7522 and 0.7805, respectively.

This indicates that FS-IV provides significant additional predictive value, particularly when paired with the core feature set (FS-I).

Impact of Adding FS-III and FS-IV Together on TSS
While FS-IV clearly contributes to improving the TSS score, adding FS-III to the combination does not seem to provide further improvement. For example:

- Combination [1, 4] (without FS-III): TSS = 0.7805
- Combination [1, 2, 4] (with FS-III): TSS = 0.7522 (a slight drop)

This slight reduction in TSS when adding FS-III suggests that FS-III does not add any meaningful value and might even introduce noise, possibly due to its sparsity or lack of real-time relevance for flare prediction.

c. Which dataset led to a better TSS score (2010 or 2020)?

Based on the results, the dataset from 2010-15 led to a significantly better TSS score than the dataset from 2020-24.

Performance Comparison:

- **2010-15 Dataset:** The best TSS score for this dataset (with feature combination [1, 4]) was 0.7805.
- **2020-24 Dataset:** The best TSS score for this dataset (with the same feature combination [1, 4]) was 0.6428.

Reason for 2010-15 dataset to perform better

1. **Class Distribution and Data Completeness:** The 2010-15 dataset contains a richer and more complete set of data entries, meaning it likely captured more diverse solar flare activity, leading to better model generalization. The 2020-24 dataset, on the other hand, seems to have fewer and possibly less complete data points, as it covers more recent solar events that may have had gaps or lacked the same level of detailed monitoring compared to the earlier period.

   More importantly, the 2010-15 dataset was the one used in previous research studies, such as the one by Bobra and Couvidat in 2015. This dataset was thoroughly curated, with a balanced distribution between solar flare events and non-events, making it ideal for training and evaluation.

   In contrast, the 2020-24 dataset may have contained fewer high-magnitude solar flares or less detailed measurements, leading to a reduced ability to accurately predict solar flares using the same feature sets.

2. **Imbalanced Class Distribution:** The class distribution in the 2020-24 dataset may have also been more imbalanced, with fewer positive solar flare events (M-class or X-class flares). When the number of positive events is small compared to negative events, it becomes much harder for the model to correctly predict rare solar flares, which significantly impacts the TSS score.

NOTE: Regarding the coding part, after confirming with the TA in the tutorial, I have implemented the confusion matrix all together on all the 10 folds together for each of the combination. Instead of choosing the average or best one, I have implemented the confusion matrix on all the k folds after confirming if it's right with the TA.

Plotting the confusion matrix for all k-folds together, rather than just the best or average one, allows us to see the variability and performance consistency across different data splits. This can help in:

1. Understanding Stability: By visualizing the results for each fold, we can check if our model consistently performs well across different subsets of data or if there is significant variation in performance.
2. Detecting Overfitting/Underfitting: Inconsistent performance across folds may indicate that the model is overfitting to certain subsets or is not generalizing well, which is important to address in model evaluation.
3. Spotting Outliers: Some folds may expose problematic or outlier data points that impact performance, and visualizing all the confusion matrices can help in identifying these issues.