BIOINFORMATICS MODELING AND SIMULATION

SECB 4313

ASSIGNMENT 2

Lecturer:

DR. AZURAH BINTI A SAMAH

Group Members:

No	Name	Matric No
1	CHANG MIN XUAN	A20EC0024
2	HANIS RAFIQAH BINTI HISHAM RAZULI	A20EC0041
3	LEE JIA YEE	A20EC0063
4	NIK SYAHDINA ZULAIKHA BINTI BADRUL HISHAM	A20EC0108

1.0 4 HYPERPARAMETER

1. **Number of Trees (n_estimators)**

   Values: 100, 500
   Justification: The number of trees affects the robustness and accuracy of the model. Increasing the number of trees generally improves performance due to reduced variance but increases computational cost. Testing 100 and 500 balances performance and efficiency.

2. **Maximum Depth (max_depth)**

   Values: 10, 20
   Justification: Controls the maximum depth of each tree. Deeper trees can capture more complex patterns but may overfit. Testing depths of 10 and 20 helps find a balance between complexity and overfitting.

3. **Minimum Samples Split (min_samples_split)**

   Values: 2, 10
   Justification: Minimum number of samples required to split an internal node. Lower values allow the model to learn more detailed patterns, while higher values can prevent overfitting. Testing 2 and 10 helps balance learning detailed patterns and generalization.

4. **Maximum Leaf Nodes (max_leaf_nodes)**

   Values: 10, 20
   Justification: Limits the number of leaf nodes in the trees. Fewer leaf nodes can simplify the model and help prevent overfitting. Testing 10 and 20 helps assess the trade-off between simplicity and model performance.

2.0 TREE DIAGRAM

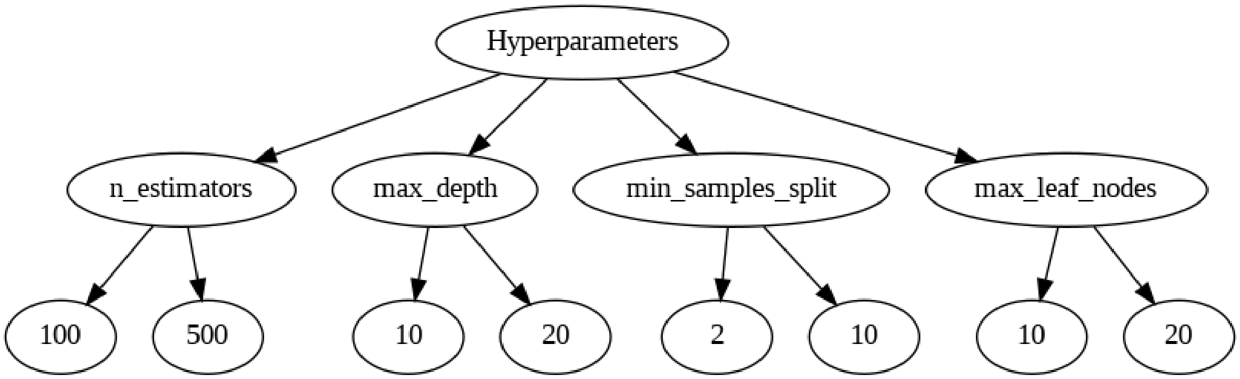

Figure 1 Tree diagram for proposed hyperparameters.

From Figure 1 above, *Hyperparameters* is the root node of the tree diagram, representing the overall set of hyperparameters being considered. *n_estimators* is a hyperparameter that likely controls the number of estimators in an ensemble model like a Random Forest. The values of 100 and 500 are shown as the potential values for this hyperparameter. *max_depth* is a hyperparameter that controls the maximum depth of the estimators in the model. The value 10 and 20 is shown as a potential value for this hyperparameter. *min_samples_split* is a hyperparameter that sets the minimum number of samples required to split a node in the decision trees or estimators. The value of 2 and 10 is shown as a potential value for this hyperparameter. *max_leaf_nodes* is a hyperparameter that controls the maximum number of leaf nodes (terminal nodes) in the decision trees or estimators. The value of 10 and 20 is shown as a potential value for this hyperparameter.

Table 1 Proposed Experimental Design

Proposed Experimental Design			
n_estimator	max_depth	min_samples_split	max_leadf_nodes
100	10	2	10
500	20	10	20

Table 1 above shows the proposed experimental design. By testing different values for these hyperparameters, we can gain insights on how the model's behavior and performance are affected by these parameters, and ultimately, we can use this information to fine-tune and optimize the model for their specific use case.

3.0 HYPERPARAMETER TUNING

Table 2 Hyperparameter tuning result

Hyperparameter Tuning Result					
Combination	**n_estimator**	**max_depth**	**min_samples_split**	**max_leaf_nodes**	**Accuracy**
1	100	10	2	10	0.8689
2	100	10	2	20	0.8525
3	100	10	10	10	0.8852
4	100	10	10	20	0.8524
5	100	20	2	10	0.8525
6	100	20	2	20	0.8852
7	100	20	10	10	0.8852
8	100	20	10	20	0.8525
9	500	10	2	10	0.8689
10	500	10	2	20	0.8689
11	500	10	10	10	0.8689
12	500	10	10	20	0.8525
13	500	20	2	10	0.8689
14	500	20	2	20	0.8689
15	500	20	10	10	0.8525
16	500	20	10	20	0.8689

Analyze results:

Based on the results in Table 2 as shown above, 3 combinations (combination 3, 6 and 7) among 16 of them generate the highest accuracy of 0.8852. The hyperparameters used are n_estimator, max_depth, min_samples_split and max_leaf_nodes. The hyperparameter n_estimator refers to the number of trees. In this case, both 100 and 500 estimators are used but the highest accuracy was achieved with 100 estimators. Increment of estimators to 500 did not improve the accuracy beyond 0.8689. The next hyperparameter, max_depth controls the maximum depth of the tree. The depth of the tree can affect the complexity of the model and deeper trees can capture more relationships in data. A max_depth of 20 generally performed well, with accuracy scores often at 0.8852. Meanwhile, a max_depth of 10 also achieved the highest accuracy with the right combination of other parameters. Furthermore, a min_samples_split parameter will evaluate the number of samples in the node. The table shows that a min_samples_split of 10 was common among the highest accuracy combinations in comparison with 2 splits. This suggests that allowing more samples to be considered for a split can help improve model performance. The last parameter, max_leaf_nodes limits the number of leaf nodes. Max_leaf_nodes of 10 were part of the highest accuracy combinations. However, setting it to 20 did not consistently improve performance, indicating that limiting the number of leaf nodes to a lower number might be beneficial. From the hyperparameter tuning results, it clearly proves that specific combinations of hyperparameters greatly influence the model performance. The top-performing combinations suggest that having a moderate number of estimators (100), deeper trees (20), and a higher threshold for splitting (10) contribute to better accuracy. Hence, the most recommended hyperparameter combination is 100 n_estimators, 20 max_depth, 10 min_samples_split and 10 max_leaf_nodes.