Discussion, Synthesis and Summary –

Using the Chocolate and Mushroom datasets, the notebooks investigate six distinct machine learning models: Random Forest, Decision Tree, Support Vector Machine (SVM), Gradient Boosting, Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN). Metrics like accuracy, precision, recall, confusion matrix, and ROC curves were used to evaluate the models' performance.

DESCRIPTION –

Depending on the dataset, the models performed at different levels. On both datasets, ensemble models Random Forest and Gradient Boosting consistently performed better than the others, obtaining the highest area under the ROC curve (AUC) and the best accuracy. Although they performed worse, simpler models like KNN and Decision Tree still produced results that were satisfactory. This implies that complex classification problems, such as those these datasets present, are especially well-suited for ensemble models. With scores above 90%, Random Forest and Gradient Boosting were the most accurate methods for the Chocolate dataset. Similar patterns were seen in the Mushroom dataset, where Random Forest and Gradient Boosting once again demonstrated superior performance. The accuracy of other models, such SVM and MLP, was marginally lower but remained competitive.

The dependability and resilience of ensemble approaches across many datasets are demonstrated by this consistency. Class imbalances were present in both datasets, which can have a detrimental effect on model performance. Techniques like undersampling and oversampling were employed to remedy this, and the outcomes made it evident that rebalancing the data enhanced the models' functionality. Specifically, false positive and negative rates were more evenly distributed, and precision and recall for minority classes were improved. This was particularly noticeable in the ROC curves, where models trained on rebalanced datasets showed better class distinction with higher AUC values. In general, the accuracy of all models was higher in the Mushroom dataset than in the Chocolate dataset.

This might be because the Mushroom dataset's more pronounced class boundaries facilitate the algorithms' ability to categorize the data. More sophisticated models like Random Forest and Gradient Boosting fared better in this situation because the Chocolate dataset, on the other hand, called for more delicate and nuanced decision-making. Because of the more pronounced class imbalance, rebalancing had a greater impact on the Chocolate dataset. Performance improved much more noticeably, especially for the minority classes. However, following rebalancing, the Mushroom dataset, which had less imbalance, displayed lower but still significant benefits. Rebalancing had an impact on both datasets, but it was particularly advantageous for datasets where the class imbalance was more noticeable.

REBALANCING THE CHOC AND MUSHROOM DATASET –

Rebalancing techniques played a critical role in boosting model performance, especially on datasets with imbalanced classes like Chocolate. Three different approaches—undersampling, oversampling, and balancing—were applied, each having a distinct impact:

- **Undersampling** – This technique, which reduces the majority class size, led to moderate accuracy for most models. Random Forest and Gradient Boosting still performed well, achieving accuracy around 85%, but simpler models like KNN and Decision Tree saw more of a performance drop due to data reduction.
- **Oversampling** – All models performed better when oversampling was applied. Even models like SVM and MLP profited from this strategy, with considerable accuracy gains, and

Random Forest and Gradient Boosting attained over 90% accuracy. By creating artificial instances for the minority class, this technique improved the models' ability to generalize.

- **Balanced Dataset** – The dataset's balance produced the best results. Accuracy levels above 92% were attained by Random Forest, Gradient Boosting, and SVM, demonstrating that a balanced class distribution improves learning and dataset generalization.

CONFUSION METRICES & MODEL COMAPRISONS –

Following rebalancing, confusion matrices and classification reports showed increases in recall and precision, particularly for the minority classes. As a result, the models performed better overall with fewer false positives and negatives. Further highlighting their capacity to effectively distinguish between classes, ROC curves also demonstrated that models trained on balanced datasets had better AUC values, with Random Forest and Gradient Boosting obtaining the highest (~0.95).

Conclusion –

The analysis emphasized how crucial rebalancing strategies are for enhancing model performance, especially when used to datasets like the Chocolate dataset that have notable class imbalances. After rebalancing, ensemble models such as Random Forest and Gradient Boosting had the greatest improvement and consistently outperformed other models. Rebalancing affected both datasets, but it affected the Chocolate dataset more than the other, showing that correcting class imbalance is essential to creating efficient classification models. Successful machine learning results on unbalanced datasets depend on both the model selection and data preparation strategies, such as rebalancing.