Discussion, Synthesis and Summary

In this assignment we explored a Drug Composition dataset which collected data to assess an individual's risk of drug use and abuse based on numerous factors. In the first step I did the preprocessing and created two separated datasets X_Choc and X_mush. There was a considerable imbalance in both the datasets which was observed after the six classification methods which were applied to both the datasets. In order to guarantee that the models could correctly capture patterns in both the majority and minority classes, rebalancing techniques including SMOTE, oversampling, and undersampling had to be applied due to the class imbalance in both datasets. We concentrated on contrasting the behaviour of the models and how rebalancing affected performance metrics including AUC, recall, accuracy, and precision

An overview of the findings and lessons learnt

The datasets X_mush and X_choc were initially trained on unbalanced data which caused biased results. The majority class was greatly favoured by most models. Consequently, the minority class's precision and recall were low.
It was observed that Random Forest and Gradient Boost showed better results than simpler models like Decision tree and KNN.

Rebalancing methods employed

All models performed much better when the datasets were rebalanced, especially for recall and F1-scores.

1. SMOTE (Synthetic Minority Over-sampling Technique): This technique significantly improved the recall of minority classes. For Instance, after using SMOTE, the recall for the minority class using SVM increased from roughly 40% to over 70%. SMOTE did, however, occasionally introduce noise, which somewhat decreased overall precision (sometimes by 5–10%).
2. Random Oversampling: This method increased the recall however it led to overfitting in decision tree models, which are susceptible to duplicate data.
   The size of the majority class was decreased through undersampling, which increased precision at the expense of overall accuracy.

Impact of rebalancing

1. Rebalancing had a significant impact in the Chocolate dataset since the data may have shown a skewed class distribution, which led to early models that overfitted to the majority class. Following rebalancing, the minority class's recall and precision probably increased, improving generalisation.

2. Depending on how classes were distributed, the Mushroom dataset may not have needed as much rebalancing. Rebalancing, however, continued to enhance performance by preventing the models from unfairly favouring one class over another.

Analysis of ROC Curves

The ability of each model to differentiate between the classes varied, according to the ROC curve study. Simpler models like Decision Trees and KNN have lower Area Under the Curve (AUC) ratings than more complex models like Random Forest and Gradient Boosting.

Lessons Learned

1. Resampling is Crucial: In order to address class imbalance, rebalancing strategies like SMOTE are essential. In the absence of them, models typically do badly on minority classes. While marginally lowering precision, SMOTE effectively increases recall. To prevent overfitting or noise addition, a balanced strategy is required.

2. It is Important to Choose Your Algorithm: Even in the absence of rebalancing, ensemble models such as Random Forest and Gradient Boosting are more adept at handling imbalanced datasets. On the other hand, less complex models such as KNN and Decision Trees need more aggressive rebalancing strategies because they are more susceptible to class imbalance.

3. Performance is Affected by Dataset Characteristics: The degree of imbalance in the dataset determines the amount of rebalancing needed.

Conclusion

The significance of resolving class imbalance in classification tasks is emphasised by this work. Rebalancing methods such as Random Oversampling, Random Undersampling, and SMOTE can be used to train models to perform better in all classes. Furthermore, selecting the right model is essential. While simpler models require more careful tweaking and resampling, more complicated models, such as Random Forest and Gradient Boosting, handle skewed data better. Rebalancing significantly increased recall and AUC in both datasets, demonstrating the importance of these methods for precise classification in imbalanced datasets.