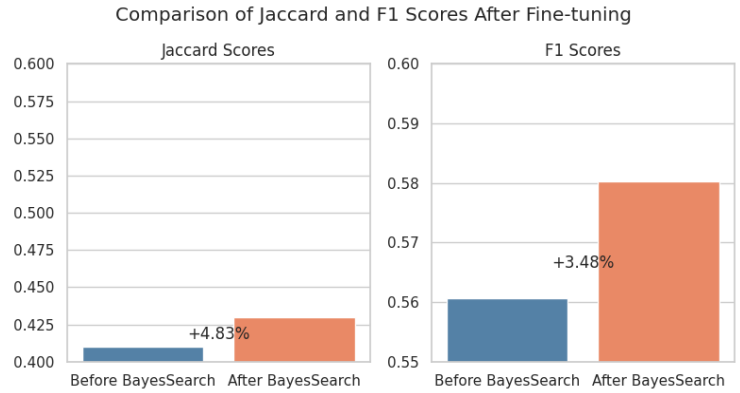
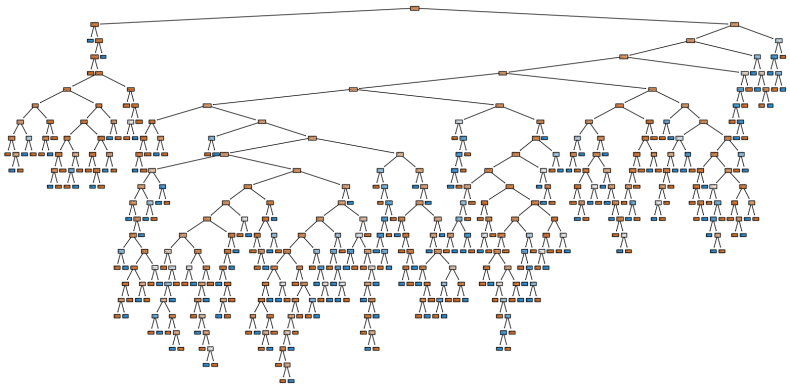
Deliverable 3

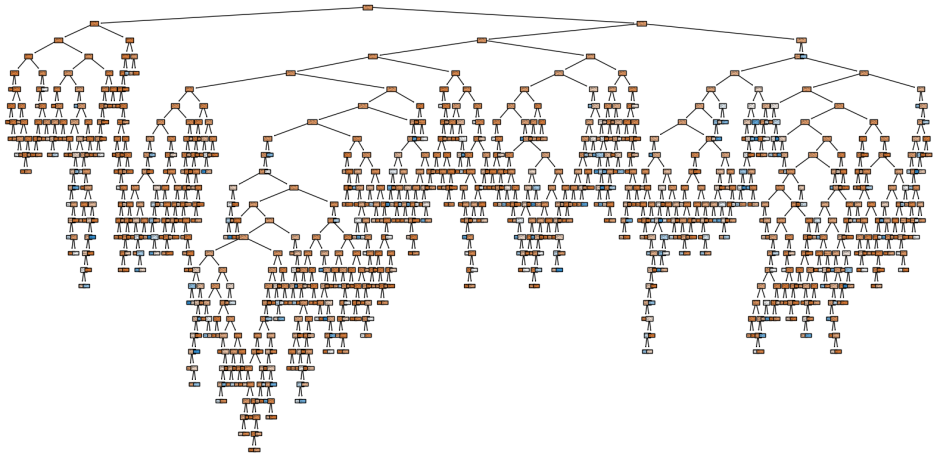
- **Final Training Results - Angela**
- The initial evaluation of the Random Forest Classifier, using the MultiOutputClassifier with specific hyperparameters, yielded Jaccard scores of 0.4100, along with F1 scores of 0.5607. After fine-tuning through BayesSearchCV, the chosen hyperparameters:
    - MultiOutputClassifier(estimator=RandomForestClassifier(bootstrap=False,
    - max_depth=37,
    - min_samples_leaf=4,
    - min_samples_split=14,
    - n_estimators=112,
    - random_state=42))
    - resulted in improved performance. The subsequent Jaccard scores increased, indicating better model accuracy (e.g., from 0.4100 to 0.4298), and the F1 scores also showed enhancement (e.g., from 0.5607 to an improved 0.5802).
- This demonstrates the efficacy of the fine-tuning process and justifies the hyperparameter choices, as the optimized model better captures underlying patterns in the data, leading to superior predictive performance.



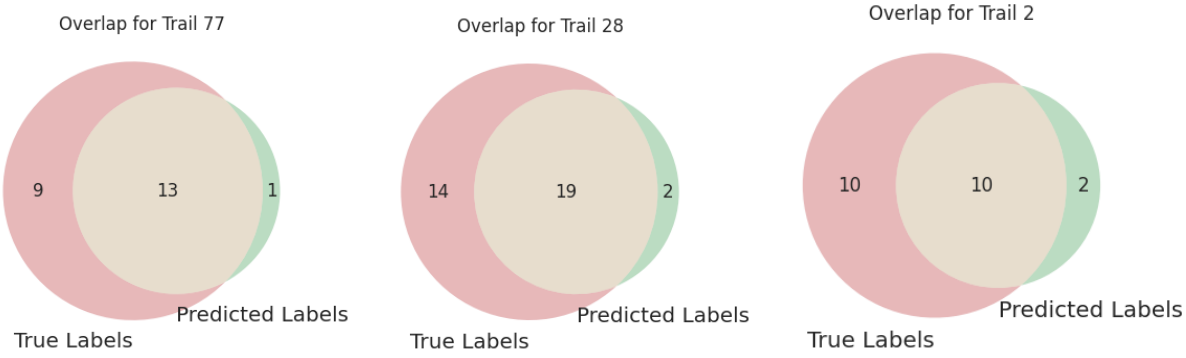
    - Before BayesSearch:



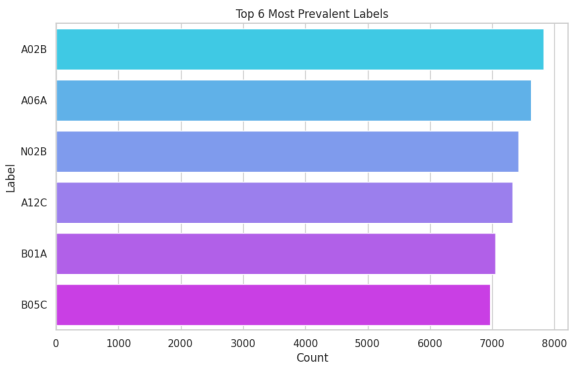
    - After BayesSearch:



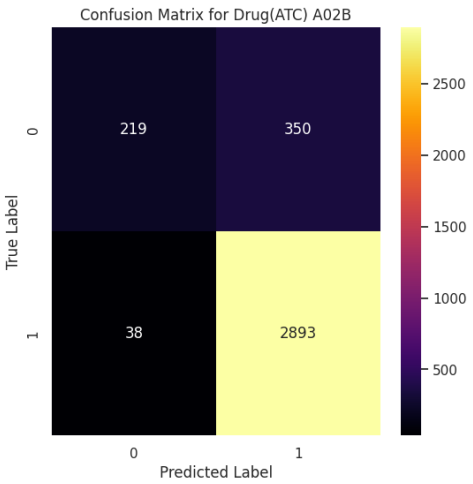
- The confusion matrix and Venn diagrams from the RandomForestClassifier's predictions on the top five common drugs showcase the model's overall strong performance, particularly evident in its high true positive rate. However, a notable concern arises with an increasing false positive rate after training on larger datasets, suggesting potential risks. In the context of drug prescription tasks, overlooking essential medications can lead to severe consequences, making it crucial to address and monitor the observed rise in false positives to maintain the model's reliability.
    - The relatively low number of false negatives underscores the significance of the model's accurate predictions in avoiding critical oversights during drug prescription. This highlights the model's efficacy in minimizing instances where crucial medications are missed, thereby reducing the risk of ineffective treatments or worsening patient conditions.
    - Continuous monitoring and refinement of the model are imperative to ensure its robust performance and mitigate the emerging issue of an elevated false positive rate, ensuring its practical applicability in real-world healthcare scenarios.
- **Venn diagrams for randomly selected trails by RandomForestClassifier**

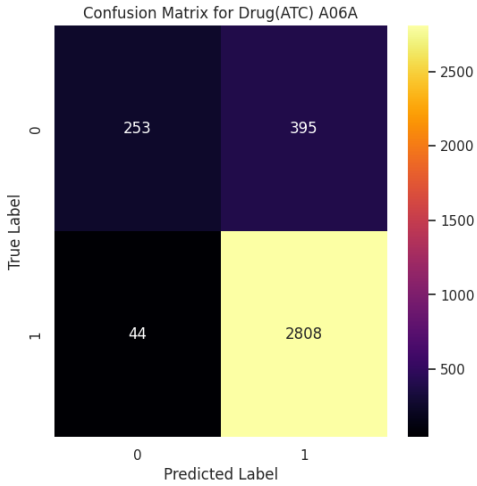


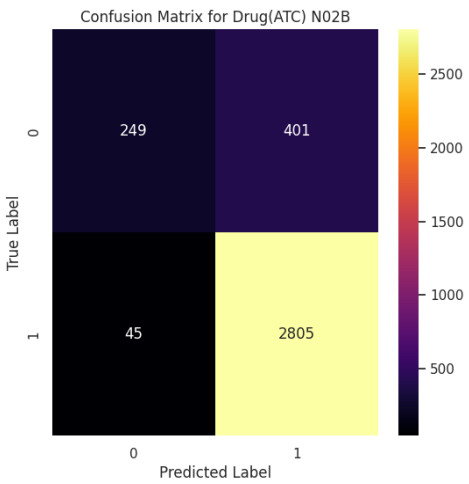
- **Top 6 most common Drugs(ATC) appeared in dataset**

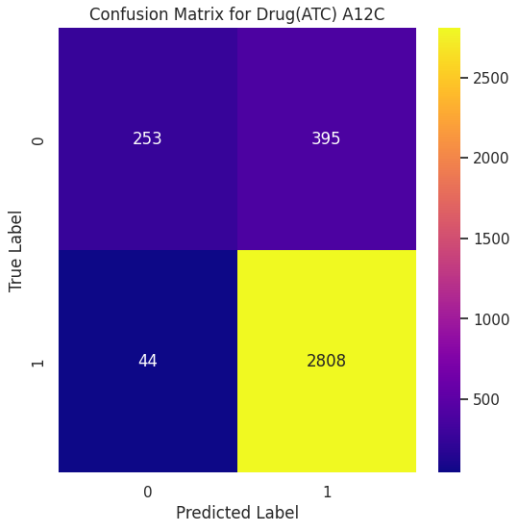


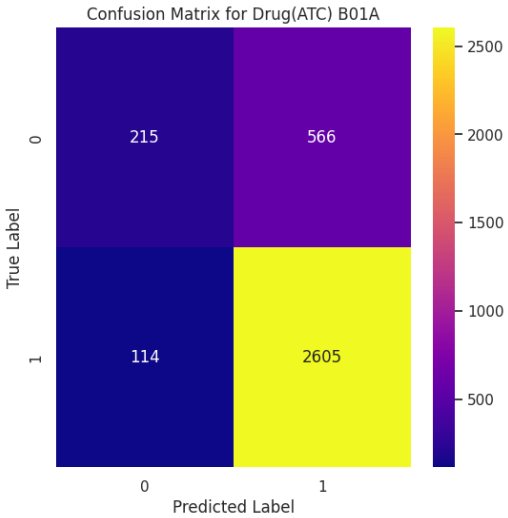
- **Confusion matrix for prediction performance by RandomForestClassifier on Top 5 most common Drugs appeared in dataset, for a total of 44204 datapoints in y_test**


Confusion Matrix for Drug(ATC) A02B
True Label
Predicted Label
0
1
219
350
38
2893
2500
2000
1500
1000
500


Confusion Matrix for Drug(ATC) A06A
True Label
Predicted Label
0
1
253
395
44
2808
2500
2000
1500
1000
500


Confusion Matrix for Drug(ATC) N02B
True Label
Predicted Label
0
1
249
401
45
2805
2500
2000
1500
1000
500


Confusion Matrix for Drug(ATC) A12C
True Label
Predicted Label
0
1
253
395
44
2808
2500
2000
1500
1000
500


Confusion Matrix for Drug(ATC) B01A
True Label
Predicted Label
0
1
215
566
114
2605
2500
2000
1500
1000
500