ML Models Uncertainty Estimation and Calibration Task

Data Preprocessing

- I identified and dropped columns with zero variability (Stddev No. Of Symbols per Categorical Features, etc.), which would be non-informative for my model.
- I used SMOTE to balance the classes, as classes were imbalanced in the dataset. I've also used train_test_split with stratification to ensure balanced class distributions in both the training and test sets.
- I explored the correlation between features but decided not to drop correlated features, as it didn't improve the model's performance. Which is reasonable because dropping features based purely on correlation might sometimes remove useful information in tree-based models.

Model Selection and Evaluation

- I've tried several models, including tree-based classifiers (XGBClassifier, RandomForestClassifier, GradientBoostingClassifier), as well as models that require scaling (LogisticRegression, SVC, KNN) and did hyperparameter tuning. Additionally, I have performed hyperparameter tuning for these models.
- Tree-based models don't need feature scaling, so standardization was applied only to the models requiring it (Logistic Regression, SVM, and KNN).
- I used accuracy (cross validation and testing) and confidence scores to evaluate each model.

Finally, I chose RandomForest_1 for achieving balanced performance across all classes. It achieved a high cross-validated accuracy of 0.8594 and the highest testing accuracy of 0.68 among the models evaluated. However, it does have a relatively low average model confidence (0.529) indicating that it may not be as confident in its predictions, meaning it needed calibration to improve its reliability and consistency.

I used three calibration methods (Platt Scaling, Isotonic Regression and Temperature Scaling)

I chose Temperature Scaling as the best choice because:

- It maintains the same accuracy (0.68) as your original model
- Has the lowest Expected Calibration Error (ECE) at 0.1282, compared to:

    Uncalibrated model: 0.1652

Platt Scaling: 0.1787

Isotonic: 0.2724

- Didn’t alter the model's predictions, only adjusts confidence scores

Platt Scaling:

- Slightly improved accuracy (0.69) but worse ECE
- Over-confident (0.84 confidence) which is unreliable given the performance
- Still fails completely on classes 0 and 4

Isotonic Regression:

- Decreased accuracy (0.61)
- Highest ECE (0.2724)
- Over-confident (0.87 confidence)
- Worse performance on several classes