1. Overview

This collection comprises datasets that describe drug side effects and indications. Originating from the Side Effect Resource (SIDER), these datasets have undergone transformations to better suit machine learning tasks. They have been further segregated into original, created, filtered, indication-specific, and results datasets. The original datasets provide an exhaustive list of side effects and indications for a range of medications. The created datasets, on the other hand, are binary representations that highlight the presence or absence of these side effects and indications for each drug. To streamline the data for machine learning tasks, certain datasets have been filtered based on the frequency of side effects or indications. There are also specific datasets created for each indication, identified by its UMLS ID.

Lastly, the results files capture the outcomes of applying different machine learning strategies, such as SMOTE/undersampling combination and SMOTE\class weight combination as well as Balanced Random forest techniques, on the datasets. These results are presented in both detailed and matrix formats, providing insights into the efficacy of these strategies in predicting drug side effects and indications.

2. File Structure and Descriptions

2.1. Original Datasets

- **meddra_all_se.tsv**: Comprehensive list of side effects for 1,430 medications. Includes STITCH compound IDs, UMLS concept IDs, MedDRA concept categories, and side effect names.
- **meddra_all_indications.tsv**: Details of 1,437 drugs and their 3,046 therapeutic roles or indications.

2.2. Created Datasets

- **Binary_Indications_data_frame.tsv**: Binary representation of indications.
- **Binary_Side_Effects_data_frame.tsv**: Binary representation of side effects.
- **Merged_Binary_data_frame.tsv**: Combined binary dataset of indications and side effects.

2.3. Filtered Datasets

- **Sorted_Indication_Count_data_frame.tsv**: Number of compounds per indication.
- **Sorted_SideEffect_Count_data_frame.tsv**: Number of compounds per side effect.
- **Filtered_Sorted_SideEffect_Count_data_frame.tsv**: Side effects with at least 10 associated compounds.
- **Filtered_Binary_SideEffects_data_frame.tsv**: Significant side effects in binary format.
- **Filtered_Sorted_Indication_Count_data_frame.tsv**: Top 10 compounds based on indications.
- **Filtered_Binary_Indications_data_set.tsv**: Most relevant indications in binary format.
- **Combined_Data_frame.tsv**: Merged dataset of filtered indications and side effects.

2.4. Datasets for Each Indication

Datasets dedicated to specific indications, identified by their UMLS IDs, such as:

- **ind_C0006826_Data_frame**
- **ind_C0009450_Data_frame**
- ... (and so on)

2.5. Results Datasets

- **The_First_results_of_SMOTE_Undersampling_***: Detailed outcomes after applying various SMOTE and undersampling strategies. These files provide insights into the performance of the Random Forest Classifier under different resampling conditions, using a 5-fold stratified cross-validation on the training data.
- **The_Matrix_result_of_***: A concise representation of the validation outcomes. It offers a streamlined view of the classifier's performance across different strategy combinations, making it easier to identify patterns and trends.
- **Max_F1_Scores_Results**: A curated list of the most effective strategy combinations. This dataset pinpoints the top-performing techniques based on the F1-score, serving as a quick reference for the best resampling strategies employed.
- **Updated_Max_F1_Scores_Results.csv**: An enhanced version of the Max F1 Scores Results. This dataset not only captures the top-performing techniques but also associates them with their corresponding original datasets (in .tsv format). It provides a comprehensive view of the strategies, their performance, and the datasets they were applied to.

- **Final_SMOTE_Undersampling_Evaluation_Results.csv**: The definitive evaluation of the applied methodologies. This dataset showcases the performance of the Random Forest Classifier using the best resampling strategies identified in the previous steps. It employs a 10-fold stratified cross-validation to assess the classifier's precision, recall, and F1-score for each dataset, providing a comprehensive view of the overall efficiency and accuracy of the methodologies on the entire data.
- **The_First_results_of_SMOTE_Classweight_{file}***: These datasets provide insights into the performance of the Random Forest Classifier when combining SMOTE resampling with different class weights. The results are based on a 5-fold stratified cross-validation on the training data. For each combination of SMOTE ratio and class weight, the average F1-score is calculated and stored. This approach allows for a detailed analysis of how class imbalance and class weights influence the classifier's performance.
- **The_Matrix_sc_result_of_***: These datasets are a matrix representation of the results obtained from combining SMOTE resampling with different class weights. Each matrix provides a clear visualization of the F1-scores across various SMOTE ratios and class weights, making it easier to identify the best performing combinations.
- **Updated_sc_Max_F1_Scores_Results.csv**: An enhanced version of the matrix results that captures the top-performing combinations of SMOTE ratios and class weights. This dataset provides a detailed view of the best strategies based on the F1-score, and it associates each strategy with its corresponding original dataset (in .tsv format).
- **Final_sc_Evaluation_Results.csv**: The definitive evaluation of the methodologies combining SMOTE resampling with different class weights. This dataset showcases the performance of the Random Forest Classifier using the best combinations identified in the previous steps. It employs a 10-fold stratified cross-validation to assess the classifier's precision, recall, and F1-score for each dataset. This dataset provides a comprehensive view of the overall efficiency and accuracy of the methodologies when considering both resampling and class weights.
- **BRF_results_{filename}.csv**: Detailed outcomes after applying the custom Balanced Random Forest classifier on each dataset. These files provide insights into the performance of the classifier under different folds of the stratified cross-validation. It captures the number of positive and negative examples in the original dataset, the number of positive and negative examples in each tree's bootstrapped sample, and the F1-score, Precision, and Recall for each fold.
- **BRF_Average_results_{filename}.csv**: This dataset provides the average results across all folds for each dataset. It offers a concise representation of the classifier's performance, making it easier to identify patterns and trends.
- **BRF_imblearn_results_**:

  This is a CSV file that captures the results of the Balanced Random Forest from imblearn library classifier for each fold of the 10-fold stratified cross-validation.For each fold.

- **BRF_imblearn_Average_results_**:

  This is another CSV file that provides the average results over all the folds of the 10-fold stratified cross-validation of the BRF imblearn.

3. **The Link to the codes:**

   - **The Balancing and algorithm:** https://colab.research.google.com/drive/1JQiWHDj4OauR-gQW15mHvikWxWko5cYy?usp=sharing

   - **The Data frame creation:**https://colab.research.google.com/drive/1YmvC8DfLfLi9PJpMINNkzA5POQ2kd-hM?usp=sharing