**Results and Discussion:**

**GloVe Results:**

**Apriori Results:**

**Machine Learning Results:**

**First Arrangement**

* The most common ADR (Pain) as labels.
* The most common disease (Hypertensive Disease) as labels.
* Coughing as the most common ADR in the AskAPatient dataset, however it will not be used in the other subsets.
* Counts for ADRs, diseases and mental issues as labels.
* Testing the best number of estimators in terms of F-Score for random forests was made by trying all estimators between 100 and 1000 at 100 increment..

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Labels/Datasets** | Hypertensive Disease | Pain | ADRCount | DiseaseCount |
| Complete | 0: 834, 1: 723 | 0: 1168, 1: 389 | 1: 695, 0: 560, 2: 302 | 1: 870, 2: 489, 0: 198 |
| Weight/Height | 1: 80, 0: 50 | 0: 89, 1: 41 | 2: 48, 1: 45, 0: 37 | 2: 64, 1: 58, 0: 8 |
| Pressure | 1: 273, 0: 188 | 0: 334, 1: 127 | 1: 229, 0: 137, 2: 95 | 1: 259, 2: 155, 0: 47 |
| Ask A Patient | 0: 608, 1: 148 | 0: 590, 1: 166 | 1: 493, 0: 194, 2: 69 | 1: 407, 0: 324, 2: 25 |

**Random Forests**

**SMOTENC versus without SMOTENC:**

In this section, the different performance metrics between the usages of synthetized dataset is compared to non-synthetized dataset.

**Hypertensive diseases:**

Hyper Tension successfully predicted with blood pressure, with weights/heights being second. This gives importance to both readings as they can predict certain diseases.

* On trials without SMOTENC, the roles were reversed, weights/heights performance increasing much higher than blood pressure, this indicates that the unmodified pressure subset is inefficient compared to weights.
* While Ask a patient’s performance fell, MedHelp complete dataset had remained stable, which is a direct result from the imbalance in ask a patient compared to MedHelp. The results however are a good indication that Hypertension can be predicted with a decent accuracy based on age alone, with other factors such as weight and height aiding the predictions.

**Pain:**

* For pain, weights/heights have the biggest effect, which could indicate that pain as an ADR is cause by weights gain in relation to height.
* On trials without SMOTENC however, all metrics drops far below the accuracy, which reflects the terrible imbalance in all the datasets, despite the fact that it is the most mentioned ADR at 25 % positives to 75% negatives in MedHelp Complete dataset.
* Ask a patient has suffered the most, with a drop from 0.7 Fscore to 0.47.

**Counts:**

* It can be observed that the number of ADRs that a person might have is much harder to predict compared to disease counts, except for ask a patient dataset, which had similar results both when using SMOTENC and without it.
* Weights/Heights are not good measures for finding possible ADRs count, with slightly better performance for discovering diseases count.
* Blood pressure was a much better indication for both ADRs and disease counts. However since disease labels are more imbalanced, the results was decreased without SMOTENC.

**Feature Filtering versus Without Feature Filtering:**

In this section, the comparison between the previous results before and after applying random forests filtering was applied, the conclusion is that it has very little effect on the results for this arrangement, this can be attributed to the use of the very limited features. This point will be further tested with arrangements 2 and 3, when all the remaining labels will be used as features.

**Hypertensive diseases:**

* Ask a patient dataset and blood pressure subset have improved their performance
* The complete dataset and weight/height subset have decrease their performance slightly, but not far off from the original none filtered features.
* With the exception of blood pressure subset, the performance on all datasets slightly decreased. Weights have dropped from 0.78 to 0.71.

**Pain:**

* Weight performances have increased, with slight decrease for the rest.
* Blood pressure and the complete MedHelp dataset improved slightly. The others decreased slightly.