Data:

The Drug Review Dataset consists of various conditions for which patients have taken a specific drug and have given review for that drug being used for a specific condition. Dataset had 885 unique conditions, to predict these many conditions is almost impossible. So, the strategy used was to create a subset dataset of top 10 most occurring conditions along with their reviews. This dataset was used for solving the problem of predicting the condition based on the reviews accurately.

Method:

There were 3 algorithms used for this purpose. Naïve Bayes (Probabilistic approach), Support Vector Machines (Hyperplane approach) and Random Forest. For Naïve Bayes, both Multinomial and Bernoulli models were used. The models were initially fed reviews as input without stemming and then to improve the model learning and performance, Snowball stemmer from NLTK was used for stemming. Stop words were removed for all the models. Both Hold out test and Cross Validation using 5-fold cross validation was carried out for all the models. For Hold out test, the data was split into 80:20 ratio as train and test data. Data is imbalanced for the target variable.

The various parameters that were tuned were Minimum Document Frequency (5, 10, 15, 20), Ngram\_range(Unigrams, Unigrams & Bigrams, Unigrams, Bigrams & Trigrams, only Bigrams, only Trigrams). For Multinomial Naïve Bayes, only Count Vectorizer with TF as input was used, Bernoulli NB Count Vectorizer with only Boolean Input. For SVM, both TFIDF Vectorizer and Count Vectorizer with Boolean, TF and TFIDF inputs were used. Random Forest was given the same input as the SVM above. For SVM, additionally the value of parameter C was tuned. For Random Forest, parameters n\_estimators (Number of Trees) and maxDepth for the trees was tuned.

Results:

For Multinomial NB, the variation of min\_df parameter from 5 to 20 in the interval of 5 was tested. As the min\_df increased, the size of the vocabulary decreased. The words less frequent in the review for a condition would be missed keeping min\_df high. For example, words like ovarian cancer, psychiatrist came only 5 times so if we were using higher min\_df we would end up losing these important words. Also, the weights were reduced when higher df was used for the common most important words. Unigrams did not help as using them gave words like make, use, think, doctor, sleep along with a limited vocab. Bigrams along with Unigrams made it better for the model to learn features like depress anxieti, suicid, antidepress. Only Bigrams usage did not contribute to the model as words such as ve use, year old were more. Unigrams & Bigrams along with min\_df 5(Vocab Size 56619) gave both the best features along with best performance for MNB. It was able to achieve precision of 0.95, recall of 0.79 and f1 score of 0.86 for ADHD condition where the row count was the least which is very good considering the Majority Vote Baseline of around 0.36 data being imbalanced.

For Bernoulli NB, where the input is Boolean, using Unigram along with Bigrams produced almost the same features but reduced model performance. Also, the model performance was increased by using min\_df of 5(Vocab Size 56619) capturing words with higher weights to contribute to the features model has learnt. The model outperforms the majority vote baseline for all the conditions to be predicted. The some of the most important features learnt by model were dermatologist, cystic, skin, acn.

For SVM, Boolean Input even though performed overall well in prediction the features learnt for model were parnat, Latuda, mania, can for Depression and for not Depression bipolar, ambien. So, using Boolean input is not a good technique for the model. TFIDF gave best features over TF input. TFIDF had most important features as diet, appetit, pound, lost, weigh for Weight Loss and for not Weight Loss condition depress, gain, hcg as for SVM we have One vs All classification. TF gave features like food energi, qysmia, 25lbs for Weight Loss and tenuat, can for not Weight Loss. So, TFIDF is the best input using TFIDF Vectorizer. Min\_df=5(Vocab Size 72390) gave the words and performance better than higher min\_df. For the choice of Ngrams\_range = (1,3) did make the most useful words. The performance was better where the model predicted conditions accurately with less errors in 7 out of 10 conditions. The value of Regularization Parameter C was tuned which changed the prediction score as well had effect on important features for a condition. The penalty was set to L1 for all models. Best value of C came out to be 0.2. The performance was better than majority vote and better than MNB.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 5-fold CV | Hold Out | Features |
| C=0.2 | 0.857 | 0.853 | Weigh,diet,lost |
| C=0.5 | 0.856 | 0.848 | Lost, fastin |
| C=0.9 | 0.854 | 0.842 | dri mouth, adipex, belviq |
| C=1.5 | 0.847 | 0.840 | Naltrexone, 13lbs, qysmia |

*Figure: SVM Parameter C tuning comparison*

Random Forest where the input was provided to be Boolean, TF, TFIDF gave the best score for Boolean Input while for TF and TFIDF it was almost less by 4%. Only for Random Forest, the best accuracy score was achieved by using highest min\_df = 20 instead of lower values. Unigrams and Bigrams gave the best overall performing model for Random Forest. The parameters specific to Random Forest algorithm which were tuned were n\_estimators 80 and maxDepth 10. Random Forest model did not perform even half as better as the least performing NB or SVM models. Random Forest showed a very low Recall and F1 score. Also, for some conditions the model was not able to predict the condition.

|  |  |  |
| --- | --- | --- |
| Model | Hold Out Test | Cross Validation (5 -fold) |
| MNB | 0.84 | 0.84 |
| BNB | 0.83 | 0.80 |
| SVM | 0.85 | 0.85 |
| Random Forest | 0.42 | 0.40 |

*Figure: Model Comparison for Hold Out and 5-fold CV*

Error Analysis on all the algorithms was done to see which conditions were mostly misclassified. Weight Loss was being wrongly predicted as Obesity and vice-versa for all models. This was because there were common words in both such as lose, weight, diet which are used for when a person wants to lose weight and for person who is suffering from obesity being linked to each other. The models were not able to distinguish this subtle difference between the two conditions. Log ratio difference between conditional probabilities for these 2 conditions gave distinct words athlet, week cardio, 225lbs which distinguishes both from each other.