Assignment3-Group5

To make the most out of this collaboration, we agreed that each team member would independently complete at least one algorithm for both the classification and clustering tasks. Afterward, we compared our individual results, analyzed the differences in our modeling processes, and learned from each other's approaches. This method ensured that we gained a comprehensive understanding of the tasks while also facilitating mutual learning.

In the following sections:

- **Data Exploration** (mainly based on the results of `JiangboSong, 23101424`)
- **Classification Task** (mainly based on the results of `YingCao, 24236217`)
- **Clustering Task** (mainly based on the results of Raja `Kiran Koushika, 24241163`)

we have supplemented the explanations with observations about our differences in approach, along with additional notes and clarifications where relevant.

Data Exploration

Class Distribution

- The dataset is highly imbalanced: **FLU**: 56.2%, **ALLERGY**: 36.8%, **COVID**: 4.6%, **COLD**: 2.3%
- This imbalance may lead classification models to favor majority classes, requiring balancing techniques to improve fairness and accuracy.

Symptom Frequency

- **Common symptoms**: **COUGH**, **SNEEZING**, and **MUSCLE_ACHES** are prevalent, occurring over 20,000 times.
- **Rare symptoms**: **PINK_EYE** and **ITCHY_MOUTH** are less frequent, occurring fewer than 7,500 times.
- Frequent symptoms often span multiple conditions, while rare ones may serve as specific markers for classification.

Feature Correlation

- Most symptoms are weakly correlated, meaning they can be treated as independent for modeling.

Moderate correlations exist within **Group 1** (e.g., **FEVER** with **NAUSEA** and **SHORTNESS_OF_BREATH**) and **Group 2** (e.g., **ITCHY_NOSE** with **ITCHY_EYES** and **PINK_EYE**). These groups are negatively correlated with each other (~ -0.33), representing distinct symptom patterns.

Relationship Between Classes and Symptoms

ALLERGY is linked to symptoms like **ITCHY_NOSE**, **ITCHY_EYES**, and **PINK_EYE**, but not all cases show these symptoms, indicating variability.

FLU and **COVID** share systemic symptoms such as **NAUSEA** and **DIFFICULTY_BREATHING**, making differentiation challenging.

Surprisingly, **COVID** lacks strong associations with sensory symptoms like **LOSS_OF_TASTE** and **LOSS_OF_SMELL**, contradicting common assumptions.

Summary of Challenges and Opportunities

Class Imbalance: The dominance of **FLU** and **ALLERGY** highlights the need for balancing methods like SMOTE or weighted loss functions.

Diverse Symptom Frequency: Highly prevalent symptoms aid classification across conditions, while rare symptoms provide specificity.

Distinct Symptom Groups: Systemic symptoms (**Group 1**) and allergy-specific symptoms (**Group 2**) can guide model interactions and clustering analysis.

Variability in Conditions: Intra-class diversity, particularly in **ALLERGY** and **COVID**, demands robust models for accurate predictions.

Classification Task

1. **Data Preprocessing**

   From the **Data Exploration** results, it was evident that the dataset was highly imbalanced.

   To address this, **RandomUnderSampler** was applied to downsample the majority classes (**FLU** and **ALLERGY**) to **7,000 samples each**, while the minority classes remained unchanged. This ensured a balanced dataset for training.

   The data was split into training and testing sets using train_test_split, with **20% allocated for testing** and **random_state=42** to ensure reproducibility.

2. **Model Training and Tuning**

Two models were trained:

- **Logistic Regression (OneVsRestClassifier)**:

  Parameters: `C=0.1 to 10, penalty='l1' or 'l2', solver='liblinear' or 'saga'`

- **SVM (Support Vector Machine):**

  Parameters: `C=0.1 to 10, kernel='linear', 'rbf' or 'poly', gamma='scale' or 'auto'`

GridSearchCV was used for hyperparameter tuning, ensuring the best parameters through cross-validation.

3. **Cross Validation**

   5-fold cross-validation was used to evaluate model performance.

   This ensured the models performed consistently across different subsets of the data, reducing dependence on a specific train-test split.

4. **Predictions & Evaluation after fitting**

   Model Performance:

   Logistic Regression: Training Accuracy = 0.9407, Validation Accuracy = 0.9260

   SVM (Support Vector Machine): Training Accuracy = 0.9410, Validation Accuracy = 0.9260

   Other detailed metrics are in the **code file**

   Both models achieved high accuracy, demonstrating:

   Effective class balancing and preprocessing.

   Strong feature discriminatory power.

Conclusion

The classification process successfully addressed class imbalance and built robust models using SVM and Logistic Regression. Both models achieved high and consistent performance, validating the effectiveness of the preprocessing, feature engineering, and parameter tuning steps.

Other:

Apart from the differences in model selection, there was also a variation in the choice of resampling methods. Some opted for SMOTE, but after comparison, we found that RandomUnderSampler is a more suitable choice for this dataset.

Raja achieved a score of 0.8755 using the Decision Tree model

Jiangbo achieved a score of 0.8794 using the Random Forest model and 0.8801 using the Decision Tree model.

Ying achieved a score of 0.9251 using the Random Forest model.

Clustering Task

Considerations in Data Preprocessing: To ensure accurate clustering and meaningful patterns, the following data preprocessing steps were applied:

Original Data: We used the original dataset instead of resampled data. Since clustering is unsupervised, it focuses on uncovering natural patterns in the data. Resampling techniques like SMOTE might introduce synthetic samples, distorting the true clustering structure. Retaining the original data ensures a more accurate reflection of symptom combinations.

No Standardization: Features were not standardized because all values are binary (0/1) and already on the same scale. Standardization would not add value in this case and could alter the inherent binary relationships.

Full Dataset: Training and validation sets were combined to maximize the dataset size, enabling the discovery of more stable and reliable clustering structures. A larger dataset also better represents the overall symptom distribution within the patient population.

Clustering Results

- **K-means**: Silhouette Score: 0.0492, Adjusted Rand Index: 0.4322
- **Gaussian Mixture Model (GMM)**: Silhouette Score: 0.0306, Adjusted Rand Index: 0.5817

Insights from Visualizations

Detailed metrics are in the **code file**

Clustering Distributions: K-means produces distinct but overlapping clusters, while GMM shows smoother transitions between groups. GMM aligns better with true labels, as reflected in its higher Adjusted Rand Index.

True Labels: Significant overlaps exist between diseases like FLU and ALLERGY due to similar symptoms.

Feature Importance: Symptoms like **Runny Nose** and **Fever** dominate PCA components, indicating they explain much of the variance in the data.

Conclusion

Both K-means and GMM revealed natural groupings in symptoms, with GMM performing slightly better. The overlap in clusters highlights the inherent challenge of symptom-based classification for diseases.

Other:

> **Jiangbo** achieved the Agglomerative Clustering with Silhouette Score: 0.0568 and Adjusted Rand Score: 0.4277
>
> **Ying** achieved Agglomerative Clustering Silhouette Score: 0.0352 and Agglomerative Clustering Adjusted Rand Index: 0.5758

Comparative Analysis and Reporting

In this analysis, we compared classification and clustering approaches. Classification, a supervised learning method, uses known labels to evaluate performance through metrics like accuracy, precision, recall, and F1-score. It provides clear and interpretable results, making it effective when labeled data is available. However, it requires labeled data, which may not always be accessible.

Clustering, on the other hand, is an unsupervised learning approach that groups similar data points without prior labels. It is evaluated using metrics such as silhouette score and Adjusted Rand Index (ARI). Clustering is valuable for uncovering hidden patterns in data but can be less interpretable and heavily dependent on the choice of algorithm and parameters.

In our findings, K-Means Clustering achieved a silhouette score of 0.0492, indicating poor separation, and an ARI of 0.4322, showing moderate alignment with true labels. Gaussian Mixture performed slightly better, with a silhouette score of 0.0306 and an ARI of 0.5817, indicating better alignment than K-Means. Both clustering methods struggled with clear separation, as reflected in the low silhouette scores. Overall, classification offers more reliable results when labels are available, while clustering serves as a useful tool for exploratory data analysis.