II. Model documentation and write-up

Information included in this section may be shared publicly with challenge results. You can respond to these questions in an e-mail or as an attached file. Please number your responses.

1. Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.
   Ifechukwu is an experienced Data Scientist in the CPG industry. His background is in engineering, and he holds a master's degree in analytics from Georgia Institute of technology.

2. What motivated you to compete in this challenge? ***I was driven by the opportunity to work with medical related data, while contributing to solving an important challenge that could positively impact the lives of many.***

3. High level summary of your approach: what did you do and why? ***My approach made use of embeddings, dimensionality reduction, clustering algorithms, network graphs and text summarization techniques to effectively identify and understand themes from medical narratives on older adults falls. The methods include: (i) Clustering: The analysis employed DBSCAN algorithm in conjunction with UMAP processing on the dimension reduced (PCA) embeddings data to uncover key themes cluster.***
   (ii) Network graphs: these were used to explore keyword pair occurrences within narratives for the different clusters, with keyword ranking via the PageRank algorithm highlighting significant terms.
   (iii) Text ranking and summarization: These methods were applied to generate summaries for narrative clusters to provide insights into key themes.

4. Do you have any useful charts, graphs, or visualizations from the process?

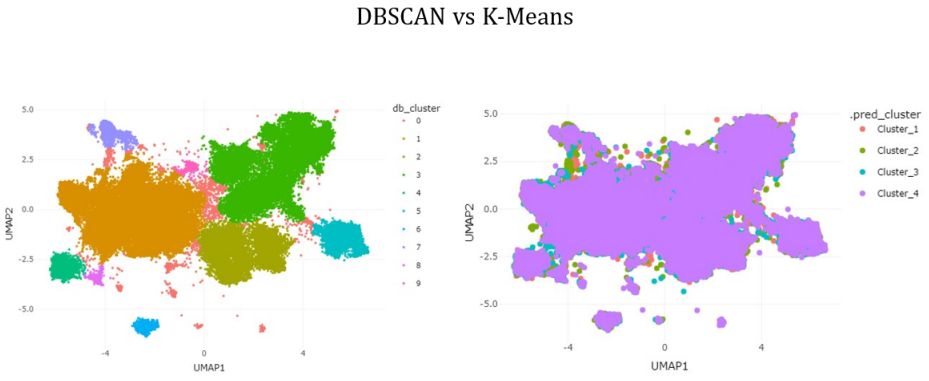

Figure 1: L-R The result of density-based clustering vs k-means

Understanding Themes

Figure 2: L-R Activities leading to falls based on theme-cluster 6; keywords associated with theme-cluster 6

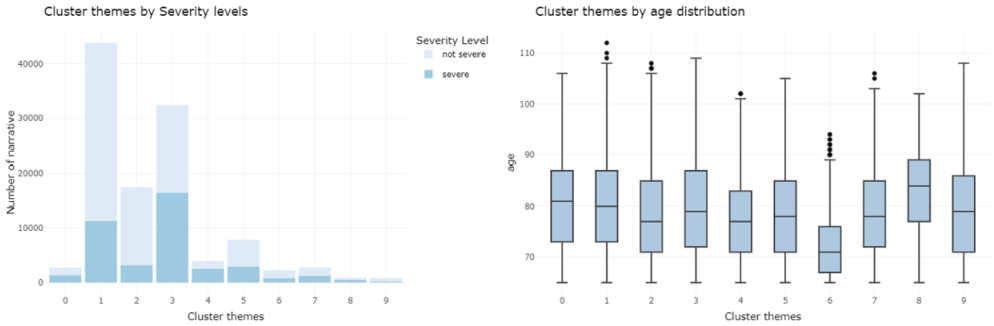

Figure 3: L-R Cluster themes by number of narratives & severity levels; cluster themes by age

5. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.

- The PCA step for dimensionality reduction

Code:

```
# Applying the dimensionality reduction step
 emb2d = recipe (~.,-cpsc_case_number, data = emb2b)|> # This line creates a recipe object using the recipe function
   step_rm(cpsc_case_number)|> # This line adds a step to the recipe to remove the variable cpsc_case_number from the dataset.
   step_normalize(all_numeric_predictors())|> # this line implements the normalization of numeric predictors
   step_pca(all_numeric_predictors()) |> # This line adds a step to the recipe to perform Principal Component Analysis (PCA) on all numeric predictors
   prep()|> # This line prepares the recipe, applying the specified steps to the data
   juice() # This line extracts the processed data from the recipe
 fwrite(emb2d, "embeddings_pca.csv") # this line copies data into a csv file called "embeddings_pca.csv"
```

- The DBSCAN step for the implementation of the clustering.

Code:

```
set.seed(123) # set seed for reproducibility
pdf_dbscan = dbscan(pdf_pca_emb_db, eps = 0.4, minPts = 550) # implementation of the dbscan algorithm on the processed data
```

- Text Summarization Step

  Code:

```
article_summary_1 <- textrank_sentences(data = article_sentences_1,
                                        terminology = article_words_1)
```

  In summary, the code is using the textrank_sentences function to generate a summary of a given set of sentences (article_sentences_1) using a specified set of terminology or keywords (article_words_1).

6. Please provide the machine specs and time you used to run your model.
   - CPU (model): **INTEL**
   - GPU (model or N/A): **NVIDIA**
   - Memory (GB): **16 GB**
   - OS: **Windows**
   - Train duration: **< 5 mins (this excludes the time for the PCA transformation)**
   - Inference duration**:< 30 secs**.

7. Anything we should watch out for or be aware of in using your notebook (e.g. code quirks, memory requirements, numerical stability issues, etc.)?
   Yes – the PCA dimensionality step is implemented on the embeddings CSV and exported to a CSV file which is read back to become the input for further processing. Due to the time it takes to implement this step, the code is run once to export the data then commented to prevent repeating the step. When uncommenting, also remove the second comment on the "step_normalize(all_numeric_predictors())" line.

8. Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?
   None

9. How did you evaluate the quality of your analysis and insights, if at all?
   I visually inspected the cluster assignment and cluster results in relation to activities based on narrative to ensure that the clusters captured the relevant activities.

10. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?
    Clustering based on word frequency did not yield good results.

11. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?
    I would have experimented with hierarchical density-based clustering algorithm. I would have also compared the cluster themes with some of the other available fields to get more insights.

12. What simplifications could be made to run your solution faster without sacrificing performance? **None**