COMP 3610: Big Data Analytics - Assignment 2

University of the West Indies, St. Augustine

Due Date: March 15th @ 11:59 PM

1 Text Analysis

The dataset for this section can be downloaded here.
It is based on reviews of 3 Disneyland branches posted by visitors on Trip Advisor.

1.1 Preprocessing, Data Organization and Visualisation (20 marks)

1. Create a new column in the dataframe called 'sentiment'. Using appropriate existing columns, populate the new column with 0's and 1's where 0 refers to a negative sentiment and 1 refers to a positive sentiment. [5 marks]
2. Clean the reviews content data and store the cleaned text in a new column 'review_content_clean'. For each step of your text cleaning give a brief explanation of why you chose to perform that method on the text. [10 marks]
3. Visualise aspects of the data to briefly summarise overall trends. [5 marks]

1.2 Text Classification (30 marks)

1. Select a metric to access the performance of your classifier and provide a brief explanation of why you chose that metric. [5 marks]
2. Perform the following classification experiments keeping track of the performance of each classification task for future use: [20 marks]
   (a) Logistic regression model on word count
   (b) Logistic regression model on TFIDF
   (c) Logistic regression model on TFIDF + ngram
   (d) Support Vector Machine model on word count
   (e) Support Vector Machine model on TFIDF
   (f) Support Vector Machine model on TFIDF + ngram

   You may use the SVM classifier from sklearn.
3. Plot a bar graph showing the performance of each of the experiments. [5 marks]

1.3 Topic Modeling (20 marks)

1. Using TFIDF and Count Vectorizer models imported for sklearn, perform topic modelling using the following topic modeling algorithms: [10 marks]
   (a) NMF
   (b) LDA
   (c) SVD
2. When choosing the number of topics give a brief explanation of why that number was chosen. [5 marks]
3. Discuss based on the top 10 words each of the algorithms choose for each topic cluster what category the topics fall under. [5 marks]

2 Classification and Clustering

The data is stored in a comma-separated file (csv) here.

2.1 Part A (10 marks)

You are required to load, explore and clean the provided dataset. Be sure to look out for values that imply the same across features. This section includes whether or not you choose to use scaling and PCA (if you use PCA, set the variance to 95%). Explain each of your steps and choices using markdown code.

Here is a breakdown of the features present in the dataset:

2.2 Part B (60 marks)

1. You will perform binary classification on the dataset to determine if a patient had a stroke or not. You are required to use 3 classifiers and compare the results using appropriate graphs and performance metrics. Your classifiers should include the Random Forest Classifier and the KNN classifier. For KNN, use cross-validation to determine an appropriate value for the number of neighbours. Use markdown code to explain your steps, choices, and results.

2. Explain the purpose of performing hyperparameter tuning.

Feature Description:

- **id**: unique identifier
- **gender**: "Male", "Female" or "Other"
- **age**: age of the patient
- **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever married**: "No" or "Yes"
- **work type**: "children", "Govt job", "Never worked", "Private" or "Self-employed"
- **Residence type**: "Rural" or "Urban"
- **avg glucose level**: average glucose level in blood
- **bmi**: body mass index
- **smoking status**: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- **stroke**: 1 if the patient had a stroke or 0 if not

3. Evaluate the performance of the machine learning models selected. Make any recommendations (minimum 2) for improvement if necessary.

2.3 Part C (20 marks)

You are required to perform clustering on the dataset using the KMeans algorithm. Your solution should include steps to find a suitable value for "k", as well as graphs showing the results. Use markdown code to explain your steps and results. The last column of the dataset is not needed in this section. State how analysing the resulting clusters (regarding the optimal cluster number only) can aid in decision making.