CS4641 B Machine Learning - Summer 2022   
Midterm report rubric

Infographic 1.0 pts   
▪ Infographic provides a concise and clear visual description of the project objectives and   
methodology and intended results   
   
Introduction 2.0 pts   
▪ Provides a clear and concise explanation of the motivation and objectives of the project

Methodology 4.0 pts   
▪ Describes the dataset in sufficient detail:   
a. Data type: image, text, numerical values, audio, miscellaneous, etc.   
b. Dataset size (# of datapoints, # of features)   
c. Targets: discrete, continuous, synthetic data, etc.   
  
▪ Describes the approach to the problem in sufficient detail:   
a. Procedure for cleaning up and standardizing the data, if applicable   
b. Procedure for feature engineering, if applicable   
c. Procedure for dimensionality reduction   
d. Unsupervised learning methods   
  
Results 3.0 pts   
▪ Presents the results of the feature selection and dimensionality reduction utilizing effective   
visualizations.   
▪ Presents the results of the unsupervised learning methods employed on the dataset utilizing   
effective visualizations.   
▪ Presents comparisons of the results of different methods applied for the same purpose   
utilizing effective visualizations   
▪ Provides visualizations for the distribution of important features and target values.   
  
Discussion 4.0 pts   
▪ Feature selection, engineering and dimensionality reduction   
a. Describes the observed relationships between different features.   
b. Explains why specific features were selected or dropped from the dataset based on the   
feature selection methods applied   
c. Provides explanation for the dimensionality reduction algorithm selected, and   
plausible justifications for the trends observed.   
d. Explains the addition of any engineered features/data augmentation, if applicable.   
   
▪ Unsupervised learning methods (density estimation, clustering, etc.)   
a. Discusses how the feature space and target variables are distributed (e.g. class   
imbalance, etc).   
b. Discusses the selection of specific evaluation metrics and why they are appropriate for   
your project objectives.   
c. Compares the performance of different methods utilized for the same purpose (e.g. K-  
Means vs. GMM).   
d. Provides explanation for differences in performance between methods based on the   
understanding of the theory and data-specific aspects.   
e. Discusses why the algorithms performed well or poorly on the dataset based on the   
understanding of the theory and data-specific aspects.

**Data preprocessing method:**

▪ Describes the approach to the problem in sufficient detail:   
a. Procedure for cleaning up and standardizing the data, if applicable   
b. Procedure for feature engineering, if applicable   
c. Procedure for dimensionality reduction

First, some features in the raw data contain string value which is difficult for machine learning algorithm to process and are converted into integer value with label encoding. For example, in the “gender” column, the “male” value is converted into 1, while the “female” value is 0.

We also observed that in our raw data, some values are missing. These missing data are filled with the mean value of the data column.

To better understand the features in the data, we plotted the correlation heat map. Features with very high correlation to each other and very low correlation to the target are subject to drop to reduce the overall dimensionality of our data.

Both PCA and T-SNE method was applied to further reduce the dimension of our data, preferably into 3D, so that the data can be visualized. We will extract the explained variance of the PCA method to understand the information we can retain after reducing the dimension.

One challenge we are facing with the raw data is that the number of data points with a positive target value is significantly larger than that of the negative value. Such unbalance may compromise our algorithm results. To rebalance the data, we used SMOTE (Synthetic Minority Oversampling TEchnique), which chose the samples with the same target value that is close in the feature space and drew new data points between these samples. We apply this method before performing the PCA and T-SNE dimensionality reduction.

**Result (Data preprocessing):**

▪ Presents the results of the feature selection and dimensionality reduction utilizing effective   
visualizations.

After the label encoding and filling in the missing data, the correlation heatmap between features and targets is plotted and shown in Figure 1. Due to the low correlation value between the “id” and the target ”stroke” we dropped this feature.

After dropping the “id” feature, we performed SMOTE to balance the data. The original data contains 4861 negative cases and only 249 positive cases. The balanced data oversample at the adjacent of the minority (positive) data points to have the same number of data points as the majority (negative) data (Figure 2).

The processed data are visualized in 3D using both PCA and T-SNE methods in Figure 3. The red X represents a positive data point, while the green dot represents a negative data point. The explained variance of different (and cumulative) principle component indexes is plotted in Figure 4.

Chart, treemap chart

Description automatically generated

Figure 1. Correlation heat map

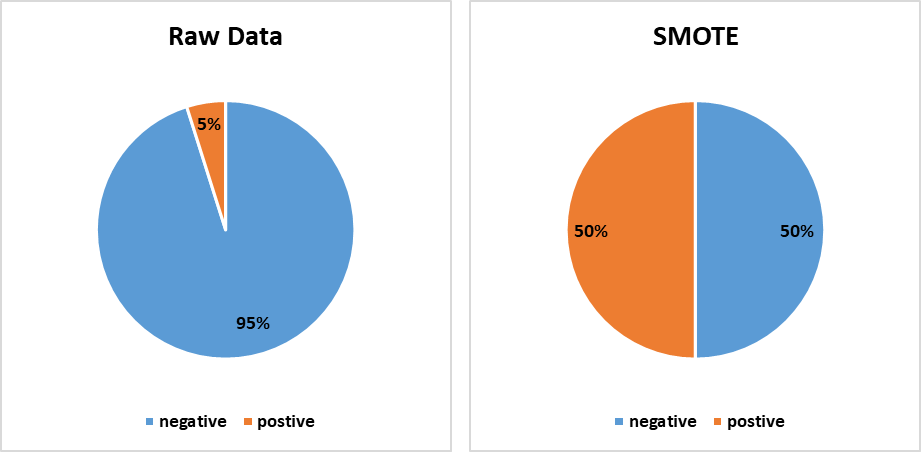


Figure 2. Using SMOTE to balance the data

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 3. 3D visualized data using PCA (left) and T-SNE(right)

(The Gif is uploaded in the same folder)

Chart, histogram

Description automatically generated

Figure 4. PCA explained variance

**Discusion Data processing**

▪ Feature selection, engineering and dimensionality reduction   
a. Describes the observed relationships between different features.   
b. Explains why specific features were selected or dropped from the dataset based on the   
feature selection methods applied   
c. Provides explanation for the dimensionality reduction algorithm selected, and   
plausible justifications for the trends observed.   
d. Explains the addition of any engineered features/data augmentation, if applicable.

From the correlation heat map (Figure 2), we can see that almost all the features are fairly independent of each other, except for “ever\_marriage” and “age” which agree with common sense. But still, the correlation between these two features is only 0.68, and we chose not to drop them. Looking at the last row, we see both the “id” and “gender” have a relatively low correlation with our target value “stroke.” The “id” stands for a random number given to each patient and intuitively is not relative to the chance of getting a stroke. Thus, this feature is dropped with confidence. Although gender also shows a statistically low correlation to target, we still decided to keep it and leave for further steps.

In the PCA explained variance, we can see that to keep over 90% of the information, we need to maintain 9 dimensions from our 11 features, which is very difficult to visualize. The 3D visualized plot, which we as humans can understand, can only explain about 40% of data variance. Therefore, in such visualized data, either with PCA or T-SNE, we see the data points with the different targets are still heavily overlapping, and unsupervised learning in 3D or lower dimension is very unlikely to distinguish the two targets.