Hope Mumme

**BMI-500 Homework 07 Report**

**Introduction**

To identify all patients who met the criteria for Acute Kidney Injury (AKI), I downloaded the clinical lab event files from the matched MIMIC-III databse from physionet. I was unable to query the matched waveform files through the MIMIC-III Postgres tutorial due to problems with my windows system (unable to connect to MIMIC-III SQL database). I also attempted to connect to the MIMIC-III SQL databse through BigQuery, but I was unable to due to not having credentialed access to physionet. As an alternative solution, I downloaded the clinical lab event file and lab classification file from the MIMIC-III Clinical Database Demo. I was able to form a local SQL database using those files and query from them. I extracted the lab events of patients and found the values before entering AKI. I used unsupervised clustering to cluster these samples into 4 groups and performed analysis to determine if these groups were able to predict if a patient would enter a stage of AKI.

**Methods**

The clinical lab category file (D_LABITEMS.csv) from the MIMIC-III Clinical Database Demo [1] was used to extract the relevant category names of the lab tests related to AKI. The clinical lab event file (LABEVENTS.csv) was used to relate lab tests to subjects and the timepoints, test values, and subject ids were extracted from this file. To retrieve the relevant information, SQL was used within jupyter notebook through the sqlite python module [2]. A SQL server was created locally, and databases were added to it and accessed throughout the analysis.

Due to limited urine output (mL/time) information in the LABEVENTS file, only creatinine information was used in AKI stage classification. The stages were classified using the information on the lecture slides, and baseline creatinine value was set as 0.85 mg/dL based on a medical reference for normal ranges of creatinine values [3]. Stage 1 AKI was set at a range from 1.23 to 1.62 mg/dL, stage 2

AKI was set at a range from 1.7 to 2.47 mg/dL, and stage3 AKI was classified as a value greater than 2.47 mg/dL. Once AKI stages were classified, stage 0 patients (patients with values less than 1.23 mg/dL) were extracted and used in the unsupervised clustering.

The data was organized into a data frame with patients as the rows, and creatinine values in the columns. Since not every patient had the same amount of column values, missing data was filled based on the average of the present data for each patient. Using the sklearn python module [4], the data frame was scaled, principle component analysis (PCA) was performed, and the resulting data was clustered using k-means unsupervised clustering. The database was accessed again to associate the patients with their future AKI stage. This was done by finding the latest subject sample with a creatinine value and finding the AKI classification for that sample. Since there are four k-means clusters (0-3) and 4 AKI stages (0-3), different true label combinations were compared against the predicted k-means clusters to determine if the clustering had any predictive value on the future stage of AKI for the patient. Accuracy scores for each permutation were calculated using the sklearn module.

**Results**

Of the samples with creatinine values in the demo database, the majority of did not meet the threshold for AKI (62%) and were classified as stage 0. The next highest category was stage 3 (Table 1). As mentioned in the methods, the dataset was adjusted due to high variability in the amount of creatinine values per subject. This is shown in the histogram in Figure 1.

| AKI Stage | Creatinine Criteria (mg/dL) | Count |
|---|---|---|
| 0 (no AKI) | creatinine content (cc) < 1.23 | 1419 |
| 1 | 1.23 <= cc <= 1.62 | 219 |
| 2 | 1.7 <= cc <= 2.47 | 227 |
| 3 | cc > 2.47 | 429 |

**Table 1. Classification of samples as AKI in the MIMIC-III Clinical Database Demo.** The count column corresponds to the number of entries in the Demo database that contained a creatinine value

corresponding to the range of mg/dL values in the criteria column. The criteria column corresponds to the stage of AKI.

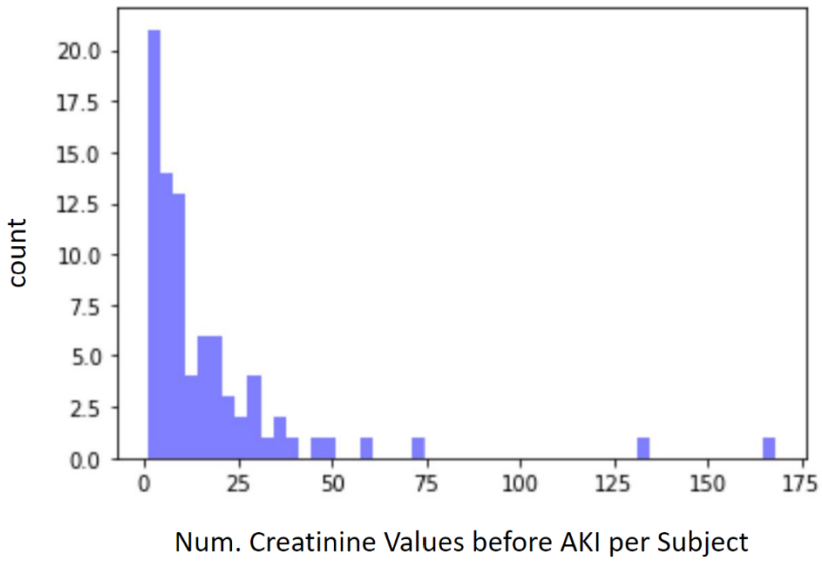


**Figure 1. Histogram of Values per Subject Counts.** The bars represent the amount of subjects that have an amount of creatinine values in that range. The values are skewed to the lower end of the x-axis.

Unsupervised clustering with k-means resulted in 4 clusters for the subject data. The feature set for this algorithm was the values of creatinine per subject that are not classified as having an AKI stage greater than or equal to one. Dimensionality reduction was performed on the scaled feature set, the top two principle components from this analysis are plotted in Figure 2A, along with the predicted k-means clusters.

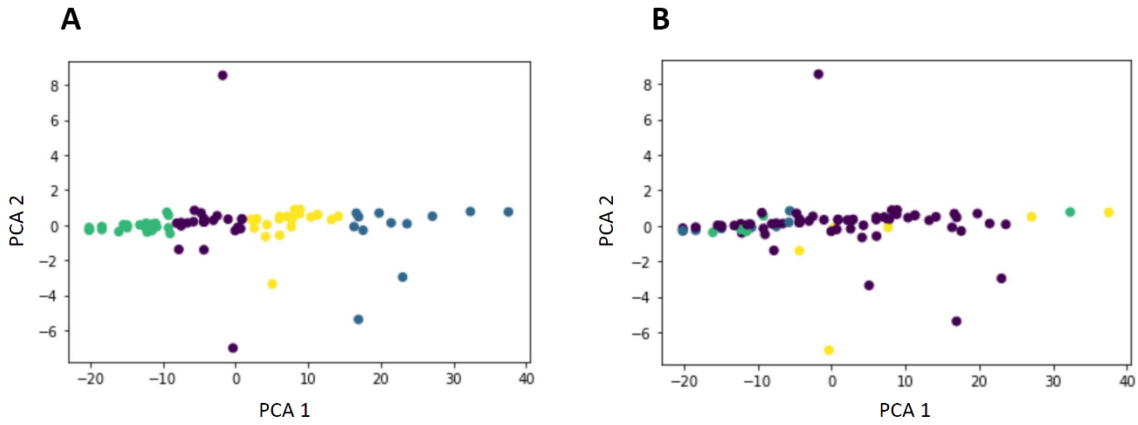

**Figure 2. PCA of pre-AKI Values with K-Means Clusters.** Principle component 1 and principle component 2 are plotted for each subject. A) K-means clusters (0-3) are represented by the colors. B) Future AKI stages (0-3) are represented by the colors.

Once the actual future AKI stages of the subject were found, they were plotted on the sample principal components, colored by the four AKI stages (0,1,2,3) (Figure 2B). The comparison of the different combinations of stages and predicted clusters were calculated and the accuracy score outputs for these comparisons are shown in the jupyter notebook on the github for this report. The highest value of the accuracy score retrieved was 0.025, with cluster 0 corresponding to stage 0, cluster 2 corresponding to stage 1, cluster 1 corresponding to stage 2, and cluster 3 corresponding to stage 3.

**Discussion**

Overall, the accuracy scores of the kmeans cluster predictions were very low, the highest being only 2.5% accuracy. This could be due to factors such as not enough PCA components were fed into the k-means algorithm, too many or too few k-means clusters were found, only creatinine values were used for classifying AKI stage, and the averages were used to fill in missing values. These factors could have affected the k-means unsupervised clustering. In addition, the PCA plots show little variability along the PCA component, this could lead to too similar k-means clusters. From Figure 3, we see that the future stages for the patients do not form distinct clusters on these two principal components, reinforcing the notion that these components are inhibiting the k-means cluster prediction.

**Conclusion**

Based on the k-means model, the pre-AKI creatinine data clusters do not predict future AKI stages. These results could be due to poor dimensionality reduction, or missing urine output information. Future steps could be to attempt other dimensionality reduction methods, using a different unsupervised clustering method such as random forest, and using a different method to fill in missing values.

**References**


1. Johnson, A., Pollard, T., & Mark, R. (2019). MIMIC-III Clinical Database Demo (version 1.4). PhysioNet. https://doi.org/10.13026/C2HM2Q.
2. Hipp, R. D. (2020). SQLite. Retrieved from https://www.sqlite.org/index.html
3. Evaluation of elevated creatinine - Differential diagnosis of symptoms | BMJ Best Practice US. (Updated September 28, 2021). Retrieved October 27, 2021, from https://bestpractice.bmj.com/topics/en-us/935
4. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
5. How to Combine PCA and K-means Clustering in Python? | 365 Data Science. (March 10, 2020). Retrieved October 27, 2021, from https://365datascience.com/tutorials/python-tutorials/pca-k-means/
6. Interacting With Your SQLite DB in a Jupyter Notebook. (2019, January 2). AP. //andrewpuleo.com/sqlite-and-jupyter/


**Code Availability**

Github repository - https://github.com/hmumme/bmi500hw07