1. Background

The dataset I chose represents the hospital records of diabetes patients in US hospitals from 1999 to 2008. Each row contains information about a given patient who underwent observations, and testing, and stayed for more than two weeks. The dataset also informs whether the patient was readmitted <30 days after being released, >30 days after being released, or NO record of re-visits [1]. This dataset is interesting because it can help identify those likely to be re-admitted after being prematurely discharged. This data can be clustered and used to predict whether a diabetes patient will have a correct discharge or a premature discharge.

2. Methods

To begin the pre-processing, I started with dropping columns that are already known to be independent of the decision to discharge a patient. Randomly generated IDs like 'encounter_id' or 'patient_nbr' fit this description. In addition, I initialized Pandas na_values=['?'], since the dataset uses '?' to indicate missing data. I use this fact to remove any rows with missing data to ensure that only patients with all known features are considered.

Next, I separate the cleaned data frame into the features we are using to cluster and the labels we are trying to predict. A problem with the features I realized was that a lot of the column data was categorical and not numeric. An example of this would be the 'gender' column, which could take on the values "male, female, and unknown/invalid' [1]. This kind of data makes it difficult to formulate a concept of distance between two points, making clustering impossible. A solution to this would be One Hot Encoding, a technique that converts discrete categories into a numeric format where all encoded columns have a 0/1 as their value [2]. Using this technique, the 'gender' column would be decomposed into 3 columns, 'Male', 'Female', and 'Other'. If a given patient were categorized as 'Male', then they would have a 1 in their Male column and a 0 in the rest. This technique would be applied to all features of d_type='object'.

Next, I scaled the data and performed K-means clustering with n=2, using the features to assign a label to a given patient. Another issue I struggled with was determining how to visualize a dataset with dozens of features into an interpretable graph. The decomposition library with Principal Component Analysis (PCA) solves this problem. PCA is an orthogonal linear transformation that allows us to project the dataset's features onto a new coordinate system. This allows us to convert our higher dimensional space into a lower dimensional space that can be represented on a graph [3].

Finally, I create a scatter plot of the clustered points using their predicted and true labels. Predicted labels are represented using colours ('pink', 'red', 'green'), while true labels are represented using symbols ('o', 'x', '+').

3. Results

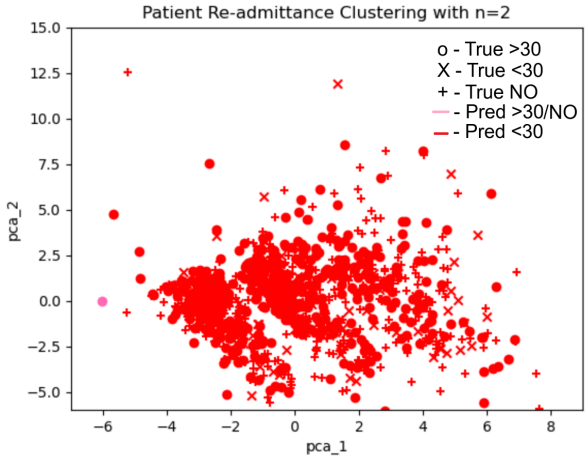

Figure 1.

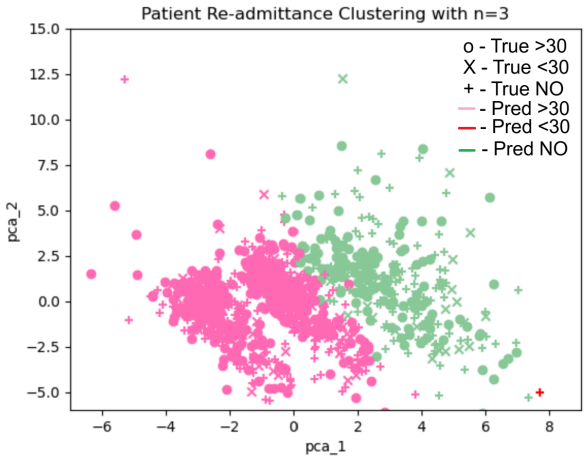

Figure 2.

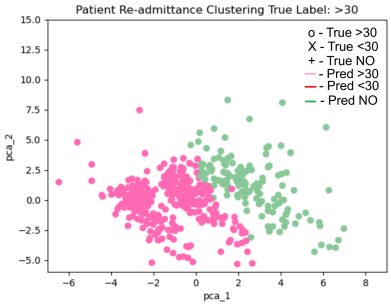

Figure 3.

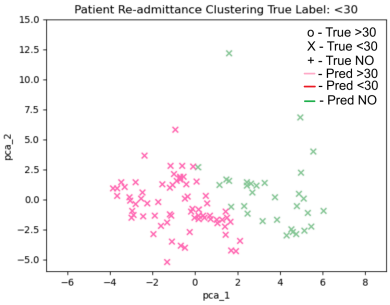

Figure 4.

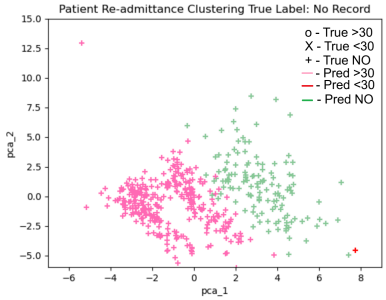

Figure 5.

4. Conclusions

Initially, I tried to cluster the data into 2 groups: premature or correct discharges. Intuitively, patients who were readmitted <30 days would be considered premature, while those who were readmitted >30 days or never readmitted would be correct. However, after performing K-Means clustering with n=2, I found that the algorithm did a poor job of predicting the true label. The 3 true groupings were poorly separated which makes it difficult to perform K-Means clustering on them (Figure 1). The clusters' non-globular shape, inconsistent densities, and outliers resulted in the creation of a single-point cluster, and a 2nd cluster that contained the rest of the points (Figure 1). This issue was resolved by re-performing K-Means clustering with n=3 (Figure 2), but many of the predicted labels still never align with their true values (Figure 3, 4, 5).

After analyzing the graphs and dataset, I can conclude that many of the features are unrelated to the likelihood of a diabetes patient being re-admitted to a hospital. A future avenue of study with this dataset could be to isolate which of these features contribute to the likelihood of being re-admitted

5. References

[1] Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). *Diabetes 130-US Hospitals for Years 1999-2008* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.

[2] Ganji, L. (2019, June 12). *One Hot Encoding in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding/

[3] *Implementing PCA in Python with scikit-learn*. (2021, February 16). GeeksforGeeks. https://www.geeksforgeeks.org/implementing-pca-in-python-with-scikit-learn/