Background

Heart failure clinical records are provided by UC Irvine Machine Learning Repository. The data collects 12 features from 299 patients during a follow up period. The patients range from 40-95 years of age, consisting of 105 women and 194 men [1, Dataset]. This data is interesting because it provides many health markers that may contribute to an accurate prediction of patients' survival during the period of time after heart failure. It is very important since 17 million people worldwide die from cardiovascular diseases [1, Background].

Some features are binary such as: anemia, diabetes, high blood pressure, sex, smoking, and death event. Binary values true and false are represented by 1 and 0 respectively. Continuous and integer attributes include age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, time. Time is the follow up period in days, meaning how many days after heart failure. Twelve of these features are used to predict the target class, death event. Death event is true if the patient died during the follow up period.

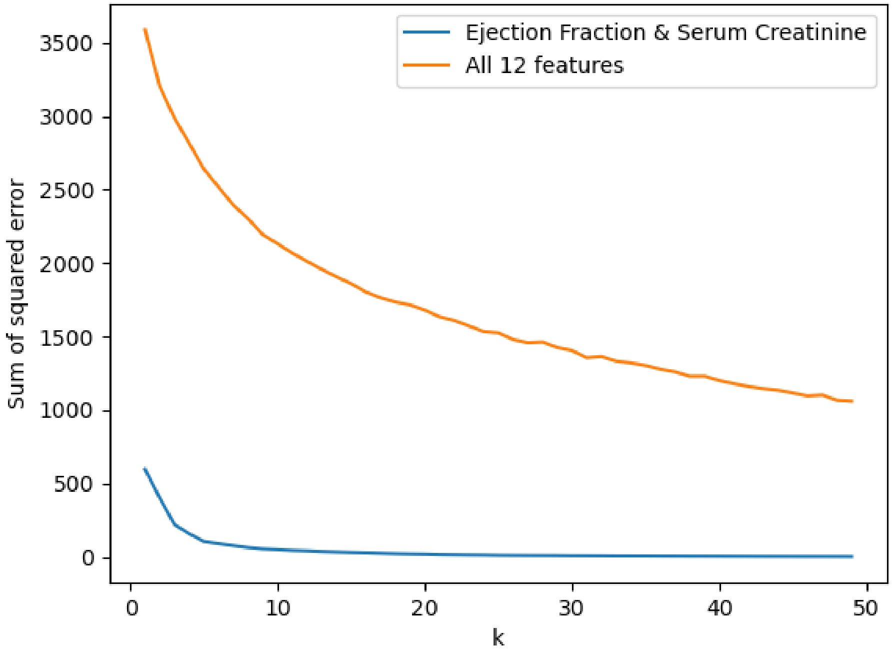
Methods

This python script attempts to group data points that are the most similar to each other into clusters. According to [1, Results], serum creatinine and ejection fraction are better predictors of heart failure survivability than using features from the whole data. With this knowledge, two methods of selecting the data were used. Below are the steps to process the data set.

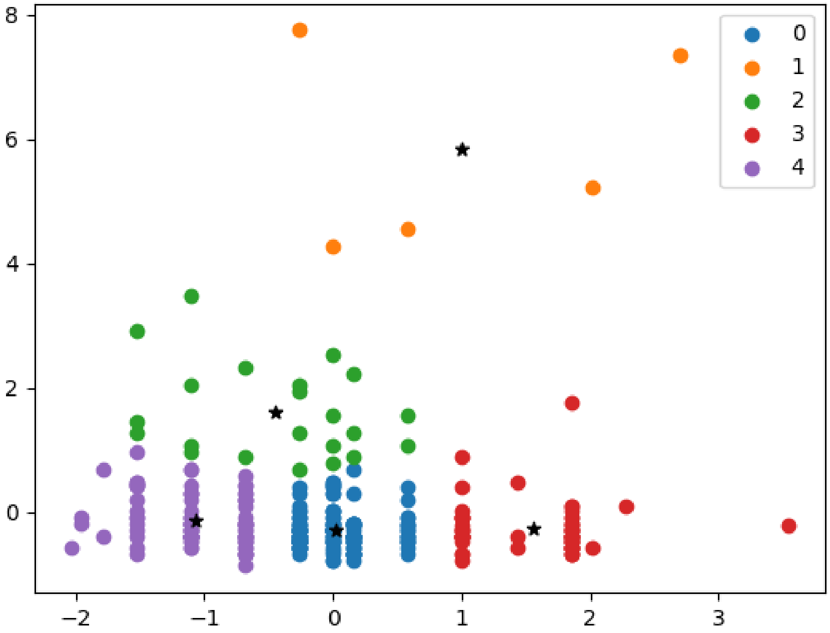
1. Load the dataset from heart_failure_clinical_records_dataset.csv file and load the necessary libraries.
2. Select features and scale the data:
    a. Method 1 used all the features except for the target class. Method 2 only used serum creatinine and ejection fraction recommended from the paper [1].
    b. Standardize and transform both methods with StandardScalar().
3. Used the elbow method to calculate the sum of squared error for both methods.
    a. Chose method 2 and its associated k value of 5.
4. Chose clustering algorithm KMeans to fit and predict the model.
5. Visualized the results with matplotlib.
    a. Plot the data and the centroids.

Results

On the graph below, method 1 which used all features of the data set had more errors for k equal 1 to 50 than method 2. In addition, there was no clear elbow for method 1 to choose a k value. Method 2 which only selected serum creatinine and ejection fraction as features had a very low sum of squared error throughout k and a clear elbow at k equal 5.



As a result, the 5 clusters below are reasonably distinct and grouped around the centroids.



Conclusions

It is confirmed that using only serum creatinine and ejection fraction produced less errors than using all 12 features. Based on the results, KMeans was an adequate method of clustering compared to other methods like Hierarchical clustering because this data did not have an obvious ordering.

References

[1] D. Chicco and G. Jurman, “Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Feb. 2020, doi: https://doi.org/10.1186/s12911-020-1023-5.