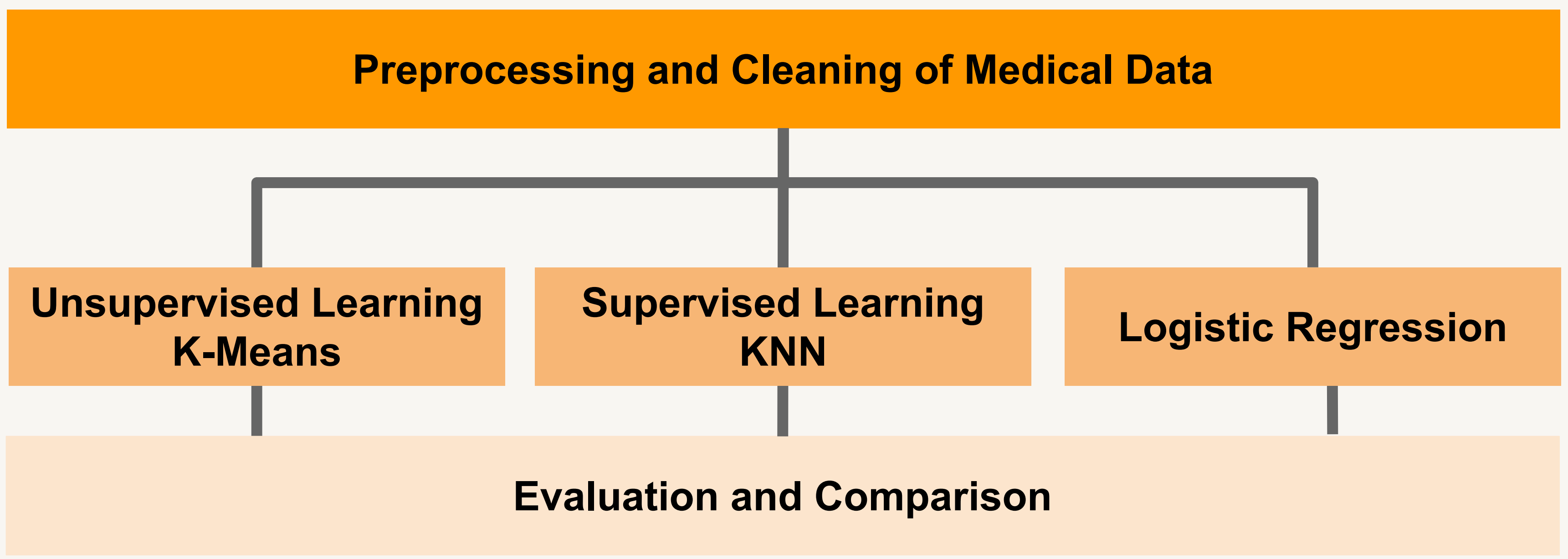
Mental Illness Diagnoses and Mortality Rate

Aileen Barry and Elizabeth Yang

Motivation

- We are interested in investigating the possible **relationship** between **mental illnesses** and **aggregate death rates.**
- **Challenges:**
    - **4.2%** of the data are **dead patients**
    - **Ambiguous data** due to **HIPAA** regulations
    - Large amount of **missing data**

Data Analysis Workflow



Preprocessing and Cleaning of Medical Data

Unsupervised Learning K-Means

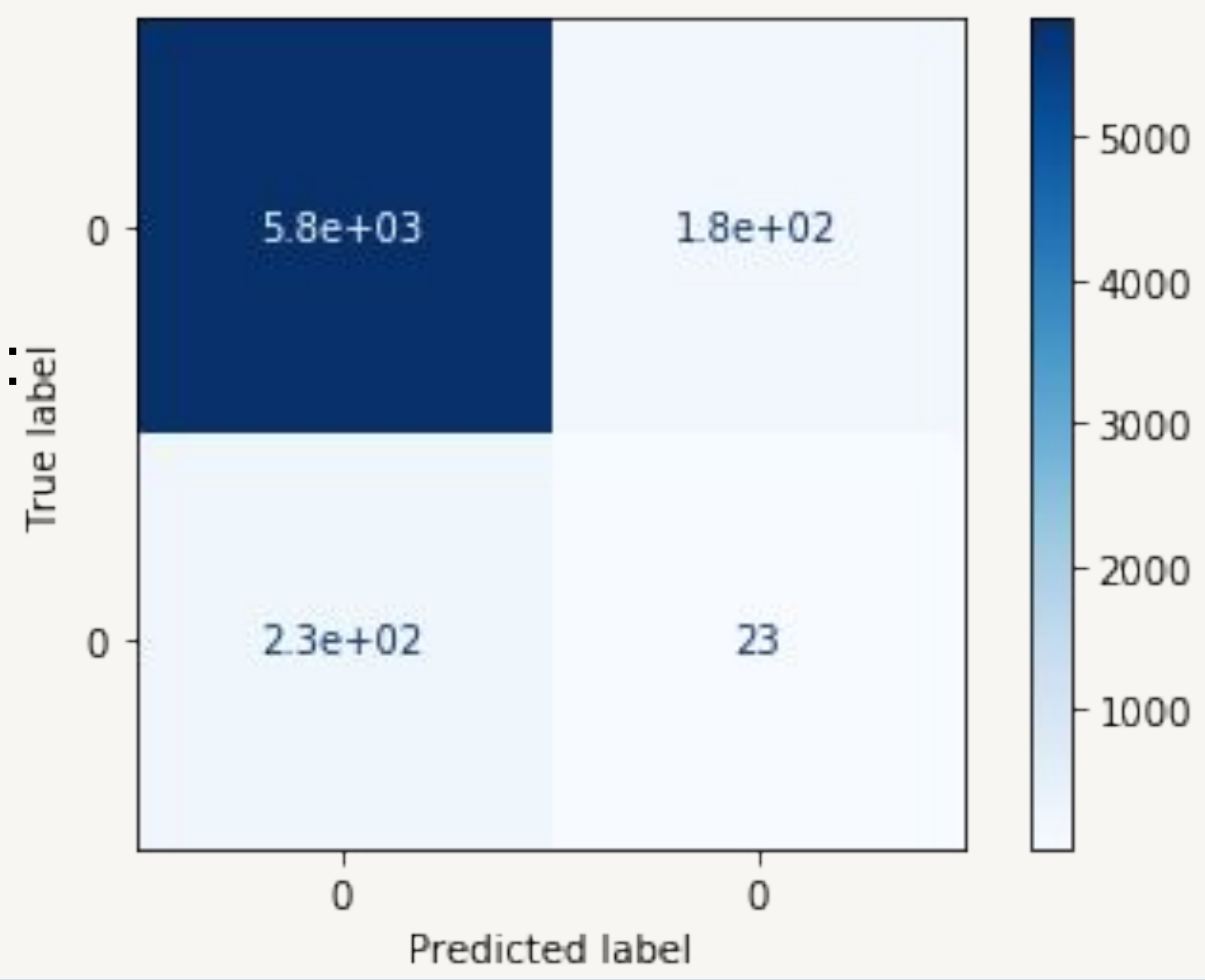
Supervised Learning KNN

Logistic Regression

Evaluation and Comparison

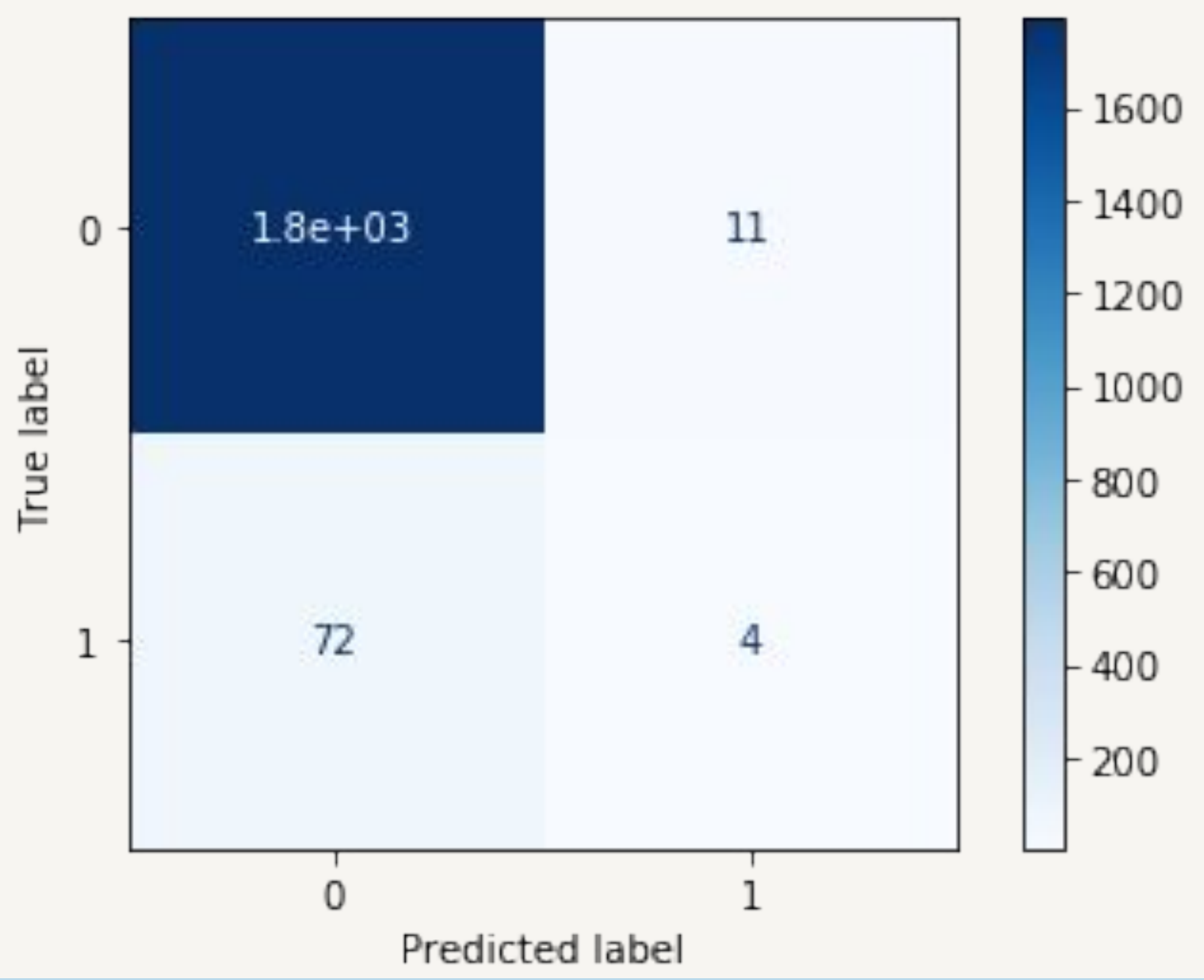
K-Means

- Pyspark implementation
- Based on hypothesis that two clusters will be divided into deceased/living patients:
- **Accuracy = 0.96**
- **Kappa = 0.07**
- (A kappa of 0 indicates that the prediction does as well as random chance)



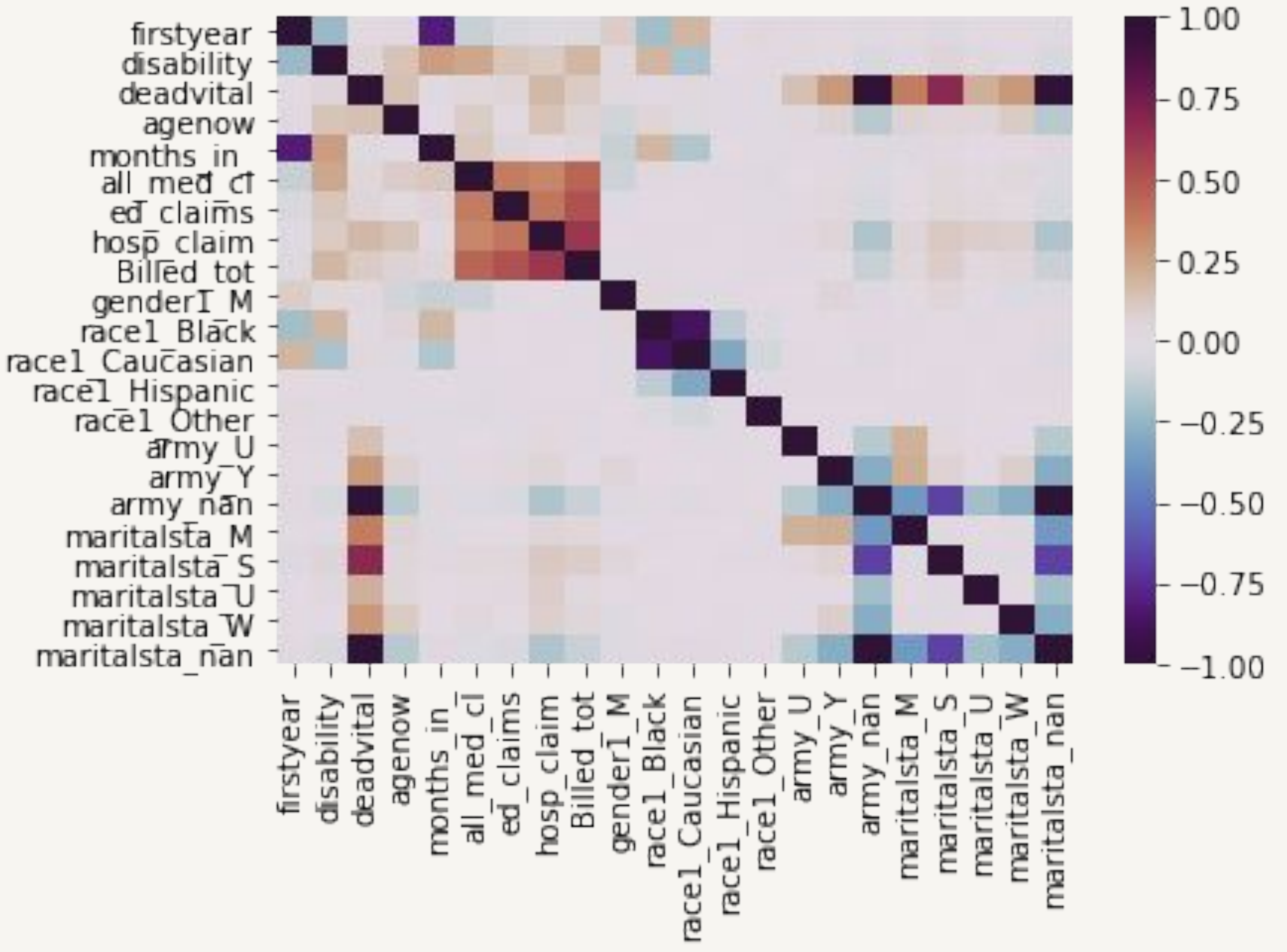
KNN

- Goal: Predict Mortality class (1 = deceased, 0 = alive)
- skLearn implementation of KNN
- 70% training/ 30% testing
- Training/testing sets are stratified to reflect unbalanced deceased/alive distribution
- **Accuracy = 0.96**
- **Kappa = 0.07**



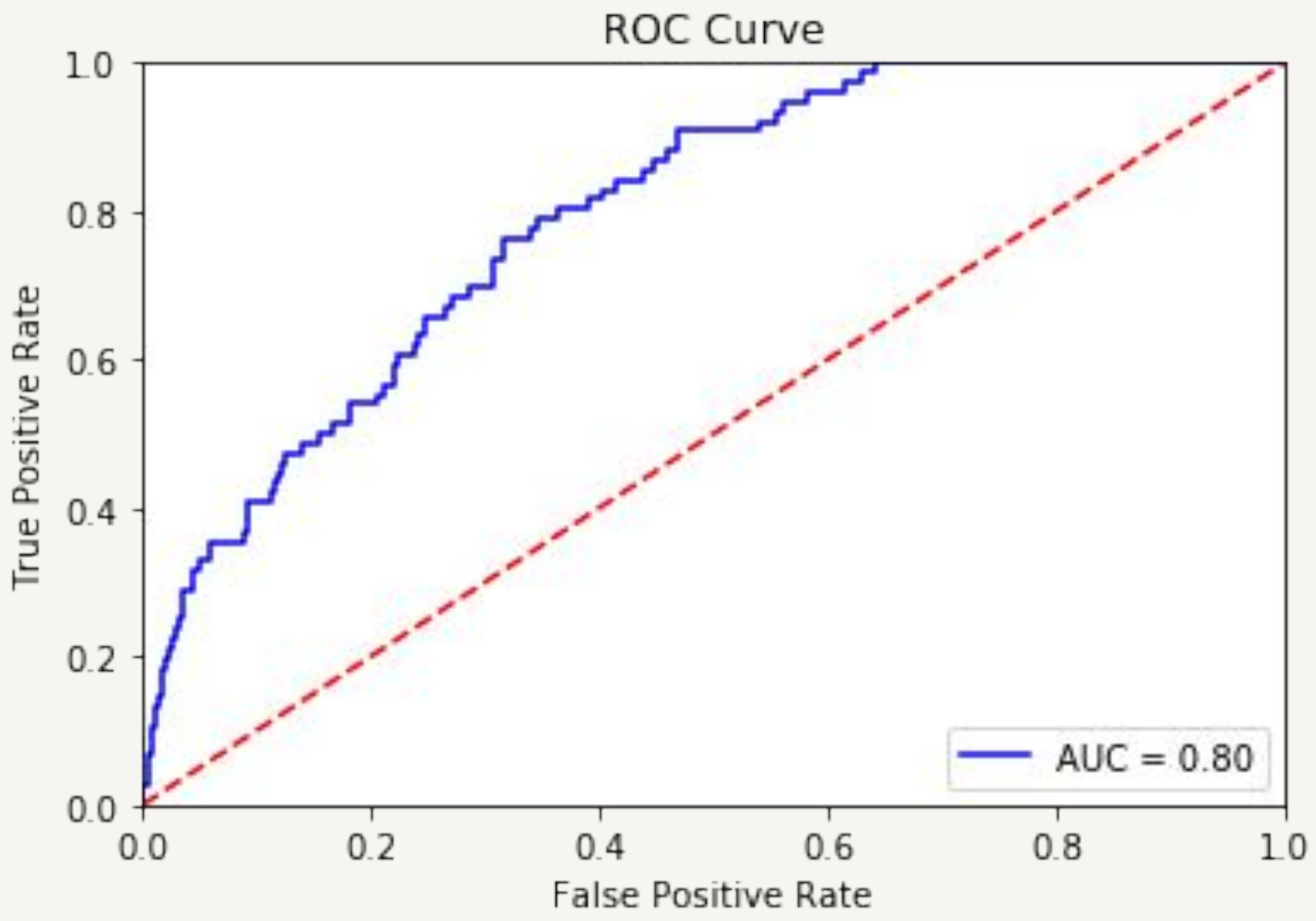
Medical Data

- **6,000** Medicaid patients in Delaware.
- **252** patient deaths recorded
- Lists each **patient's cause of death, demographic information, and Medicaid usage information.**
- Variables recorded once a patient was deceased: cause of death, year of death, type of death facility, autopsy status, military status, and marriage status.
- **Scaled numerical data** on 0 to 1 scale as most variables are binary.
- Made **dummy variables** for categorical variables



Logistic Regression

- Predict Mortality Class (1 = deceased, 0 = alive)
- skLearn implementation
- Predicted probability that individual belongs to positive class, then individual is assigned class based on threshold value
- ROC plot True/False positive rates over thresholds from 0 to 1.
- **AUC = 0.80**
- **Accuracy = 0.96**
- **Kappa = 0.07**



Evaluation and Conclusions

- Our **accuracy rates of 0.96** were inflated due to the large proportion of living patients and high probability of all algorithms to classify an individual as living.
- We used Cohen-Kappa metric to compare these success rates to random chance. Our **Kappa metrics of 0.07 are near 0**, indicating that our models did not predict much better than random chance.
- In order to benefit from our methods, there is a **need for a larger, more balanced dataset** including a larger population of deceased patients.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Citations:

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.