Machine Learning in Biology:

Predicting Heart Failure based on Patient Medical History and Blood Chemistry

Prepared by: Joseph Matthew R. Azanza

Problems Encountered

- Public datasets are too complex for traditional ML models
- Small sample sizes (tens to hundreds)
- Need to collaborate with scientists

Learning Goals

- How to handle public datasets with small sample sizes
- How ML models can be interpreted in a biological context

Dataset

- Obtained from Ahmad, et al. (2017):
  Survival analysis of heart failure patients: A case study

- 299 Pakistani patients with 12 variables, predicting mortality (299x13)

- Patient info: *age*, *sex*, *smoker*, *time*
  Medical history: *anemic*, *blood pressure*, *diabetic*
  Blood chemistry: *creatinine phosphokinase*, *blood creatinine*, *blood sodium*, *ejection fraction*, *platelets*,

Performed Methodology

Data Processing and EDA

Classification via Train-test split (8 models)

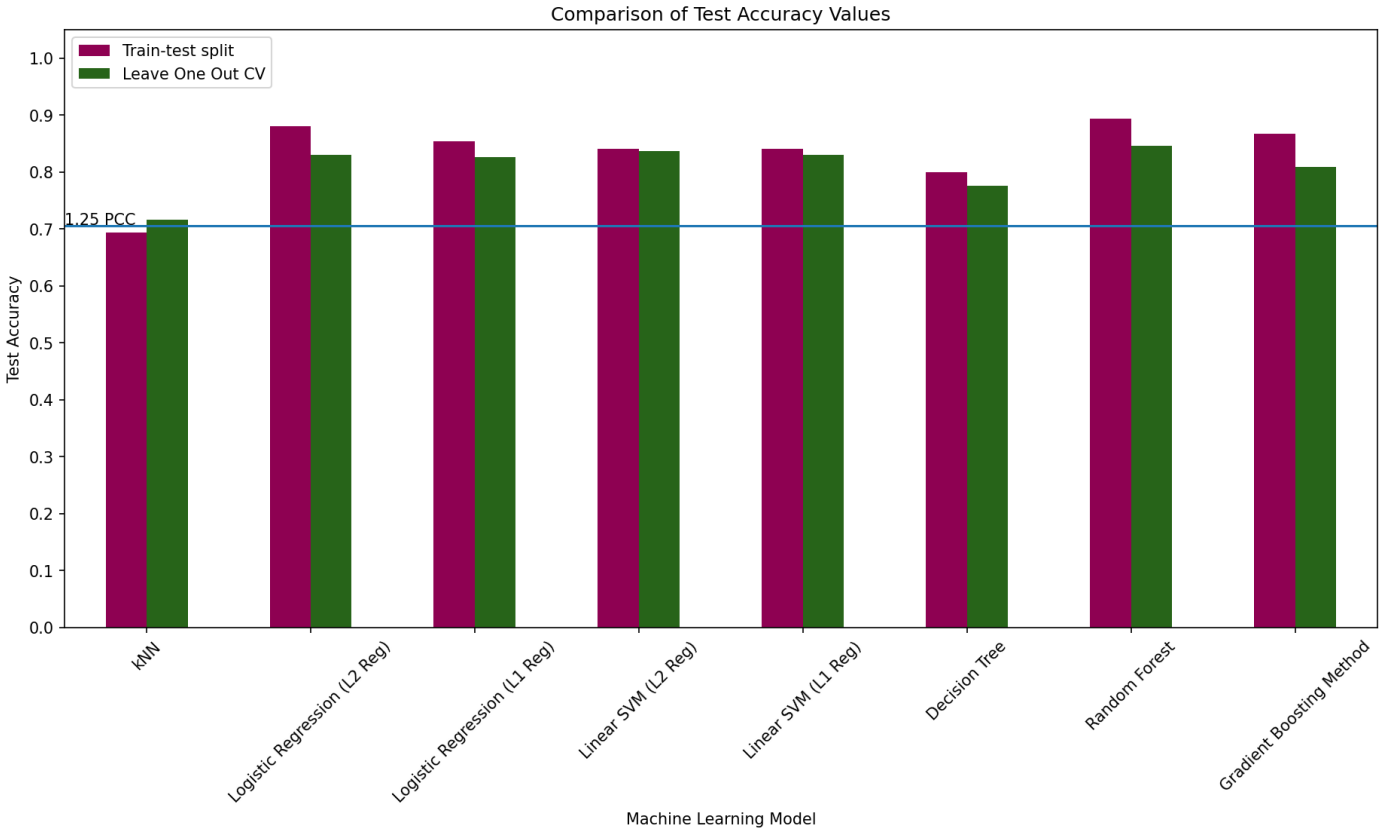
Classification Leave One Out CV (8 models)

Compare and Interpret

Leave One Out Cross Validation

- Extreme version of k-fold cross validation using all but one data point as train set; left out = test set
- Advantage: more robust model performance and no randomness
  Disadvantage: variability of model performance and computation cost
- We can use LOOCV to handle small datasets

Results


Comparison of Test Accuracy Values
Train-test split
Leave One Out CV
1.25 PCC
Test Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
kNN
Logistic Regression (L2 Reg)
Logistic Regression (L1 Reg)
Linear SVM (L2 Reg)
Linear SVM (L1 Reg)
Decision Tree
Random Forest
Gradient Boosting Method
Machine Learning Model

Interpreting ML in a Biological Problem

- Best Model: Random Forest Classifier, 85% Accuracy
- Skepticism on applying model to non-Pakistani People
- Value: ML model used to augment decision making

Summary

- We can use the Leave One Out Cross Validation method to handle small datasets
- Application and interpretation of ML in a biological problem is highly context dependent
- The value ML gives is the augmentation of decision making

Appendix

LOOCV Calculated Variance

	Machine Learning Model	Test Accuracy	Variance
0	kNN	0.715719	0.203465
1	Logistic Regression (L2 Reg)	0.829431	0.141475
2	Logistic Regression (L1 Reg)	0.826087	0.143667
3	Linear SVM (L2 Reg)	0.836120	0.137023
4	Linear SVM (L1 Reg)	0.829431	0.141475
5	Decision Tree	0.772575	0.175703
6	Random Forest	0.849498	0.127851
7	Gradient Boosting Method	0.809365	0.154294