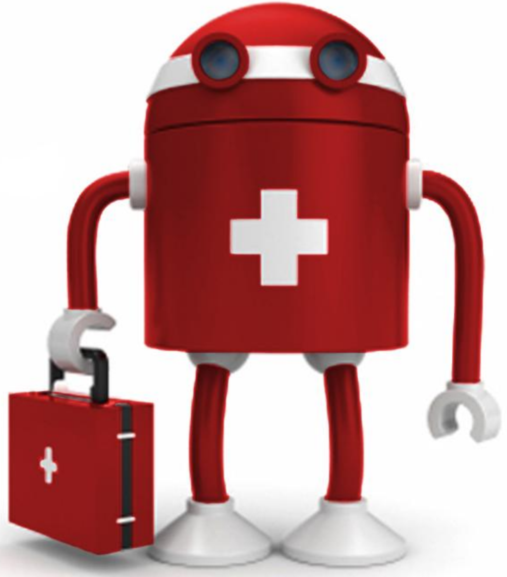


Predicting Diabetes Diagnosis in African Americans Using Ensemble Machine Learning

Kimberlyn Bailey, Jarvis Miller and Valerie Santiago-Gonzalez
Under the mentorship of Dr. Sherri Rose and
doctoral candidate Savannah Bergquist

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Objective of Our Work

Build an ensemble machine learning algorithm that can accurately predict type II diabetes diagnoses from datasets of relevant patient characteristics.

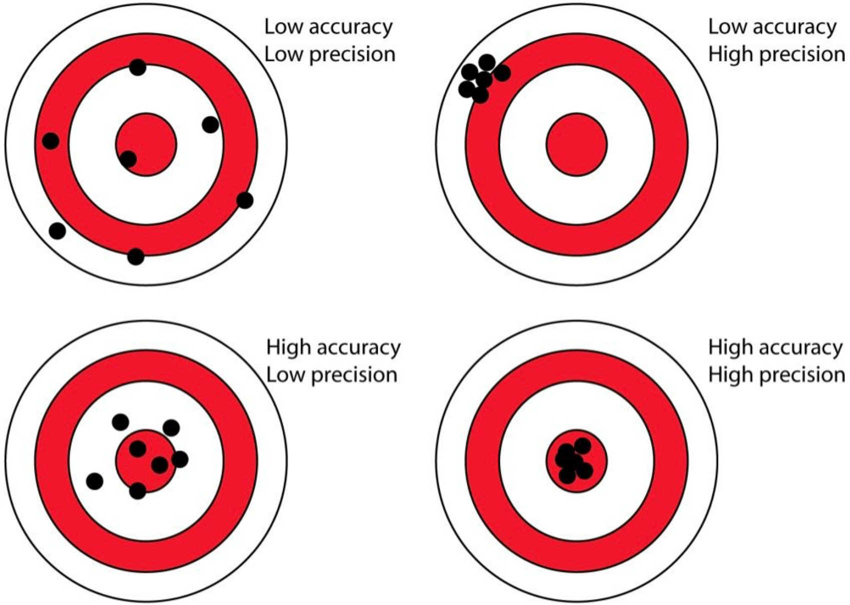
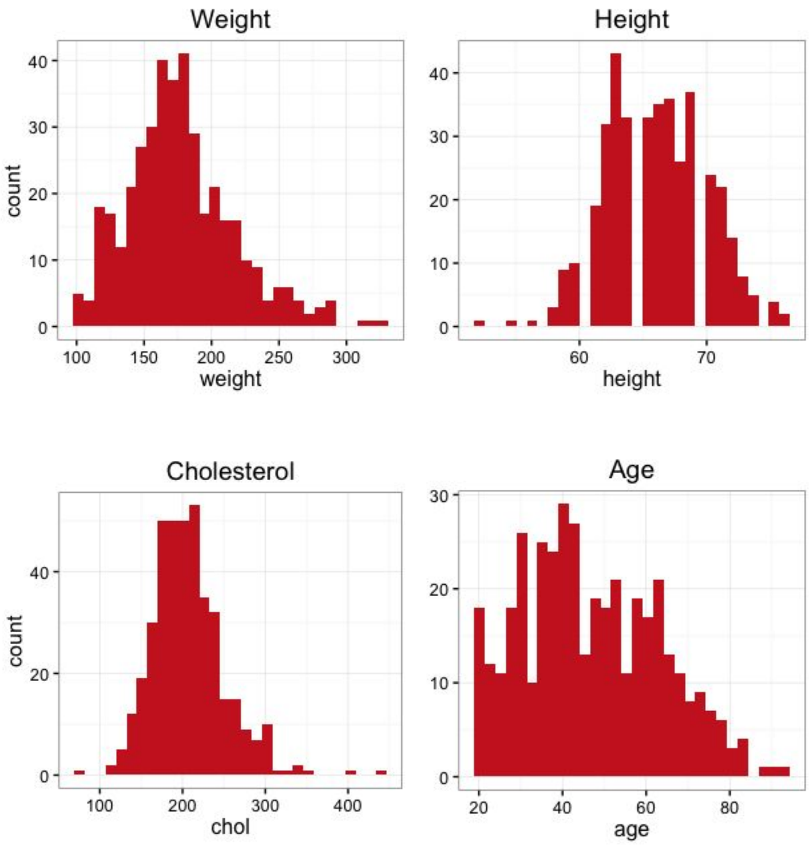


Image “Precision and accuracy in glacial geology” from antarticglaciers.org

Diabetes Dataset

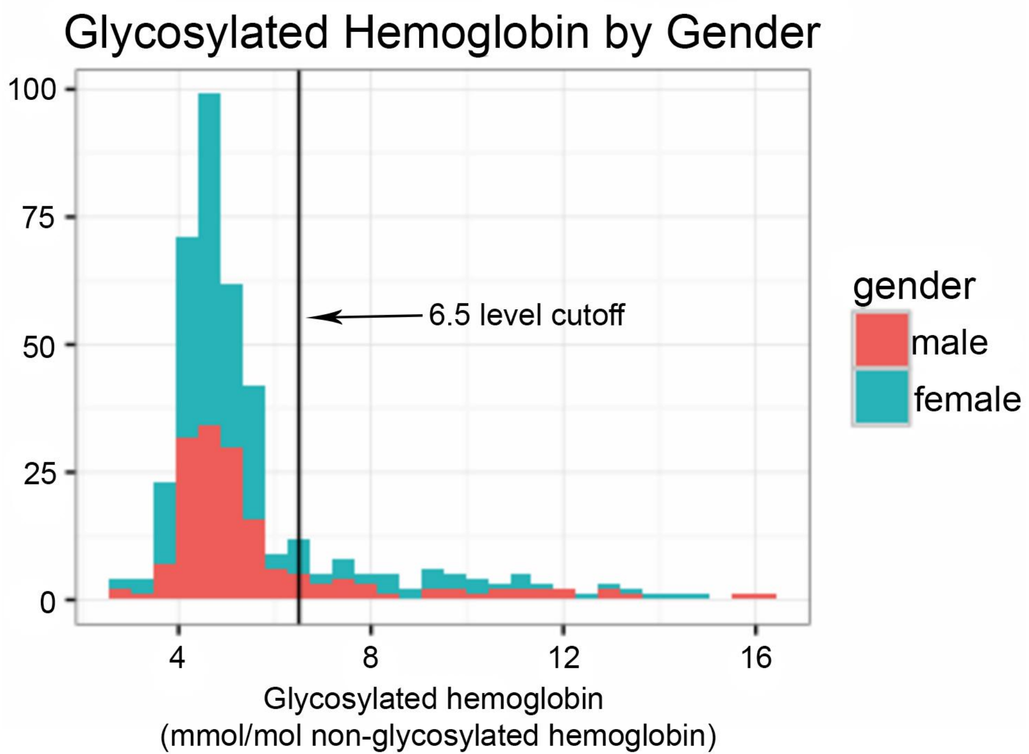
Background:

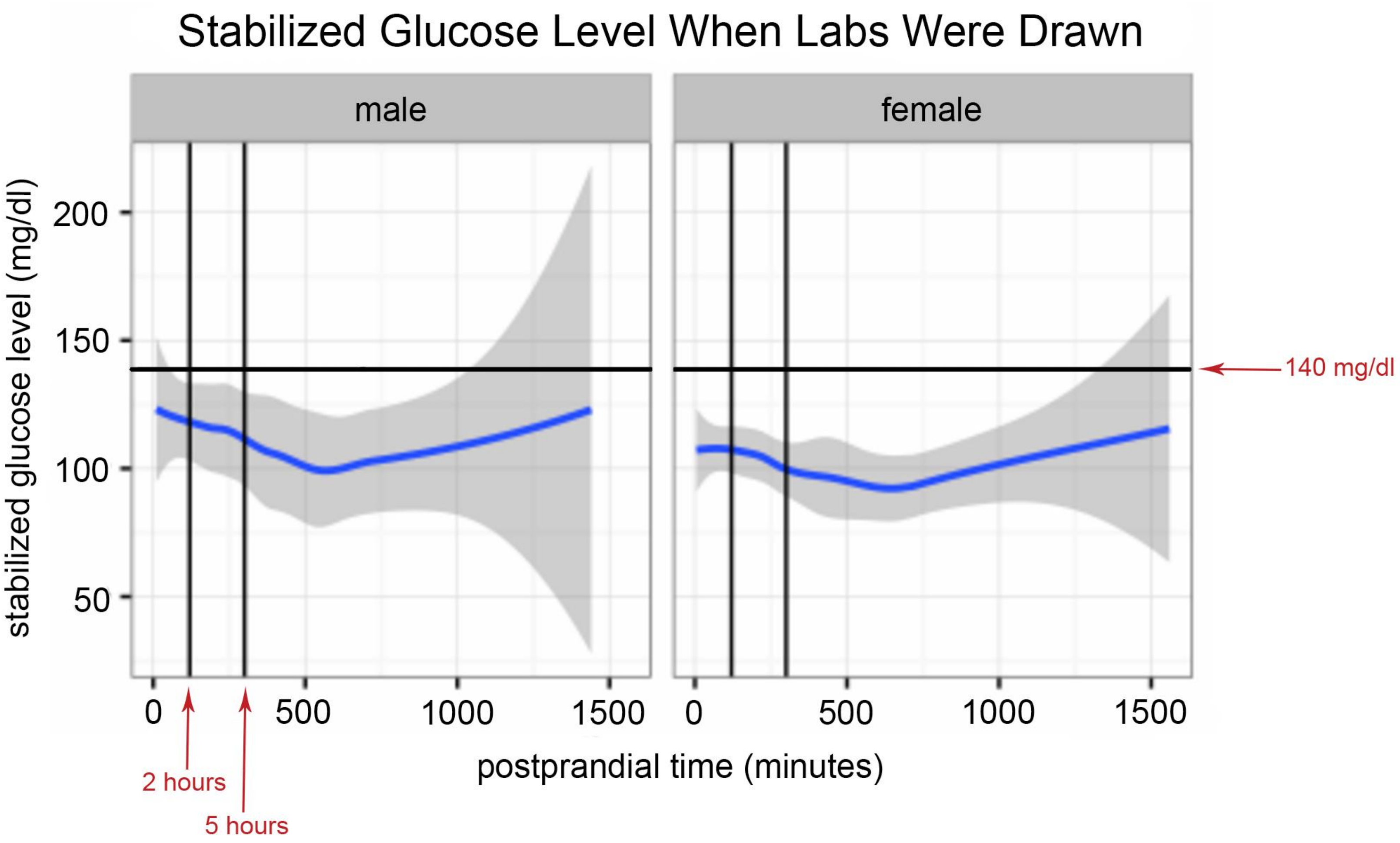
- Publically available dataset in R compiled by University of Virginia School of Medicine.
- Study designed to examine cardiovascular disease and diabetes trends of central Virginian African Americans.
- 403 subjects
- 14 variables



Data Transformation

- Missing data deleted → trimmed to 366 subjects
- Calculated *waist/hip* ratio
- Changed location variable to binary
- Created new variable: *bad cholesterol = total cholesterol - high density lipoprotein*
- Changed glycosylated hemoglobin to binary with 6.5 cutoff




Stabilized Glucose Level When Labs Were Drawn
male
female
stabilized glucose level (mg/dl)
200
150
100
50
0
500
1000
1500
postprandial time (minutes)
2 hours
5 hours
140 mg/dl

Machine Learning

Algorithms that can independently adapt or “learn” from new data.

“Machine Learning: What It Is and Why It Matters,” SAS.com
Image “Stock Robot Header” from latd.com

Facebook Newsfeed is Machine Learning

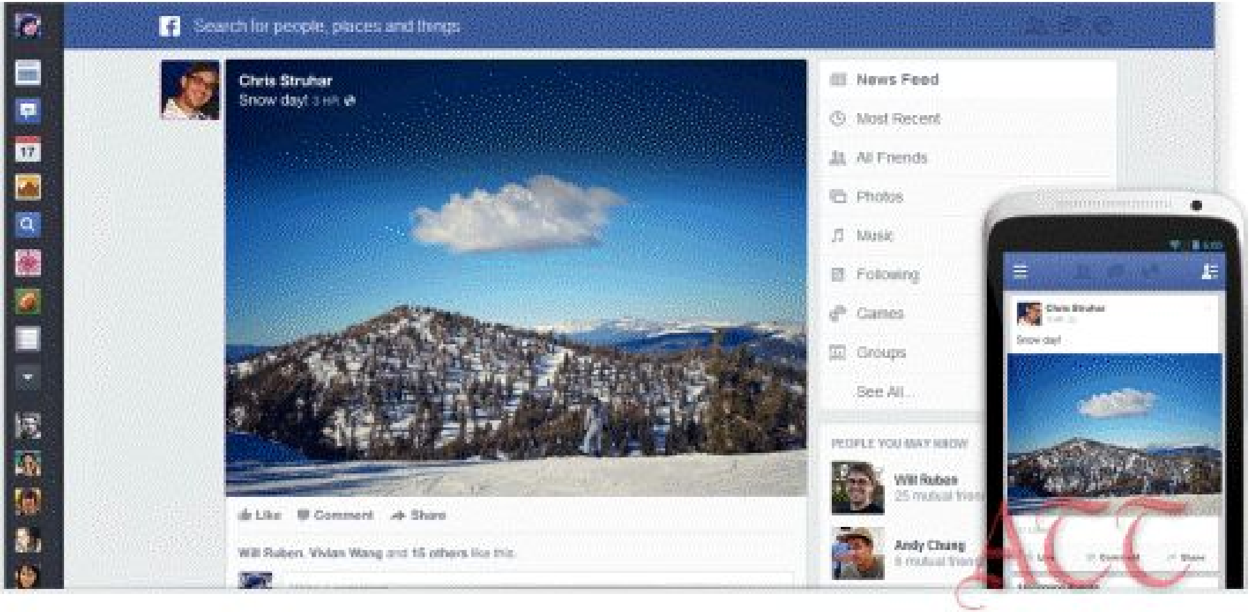
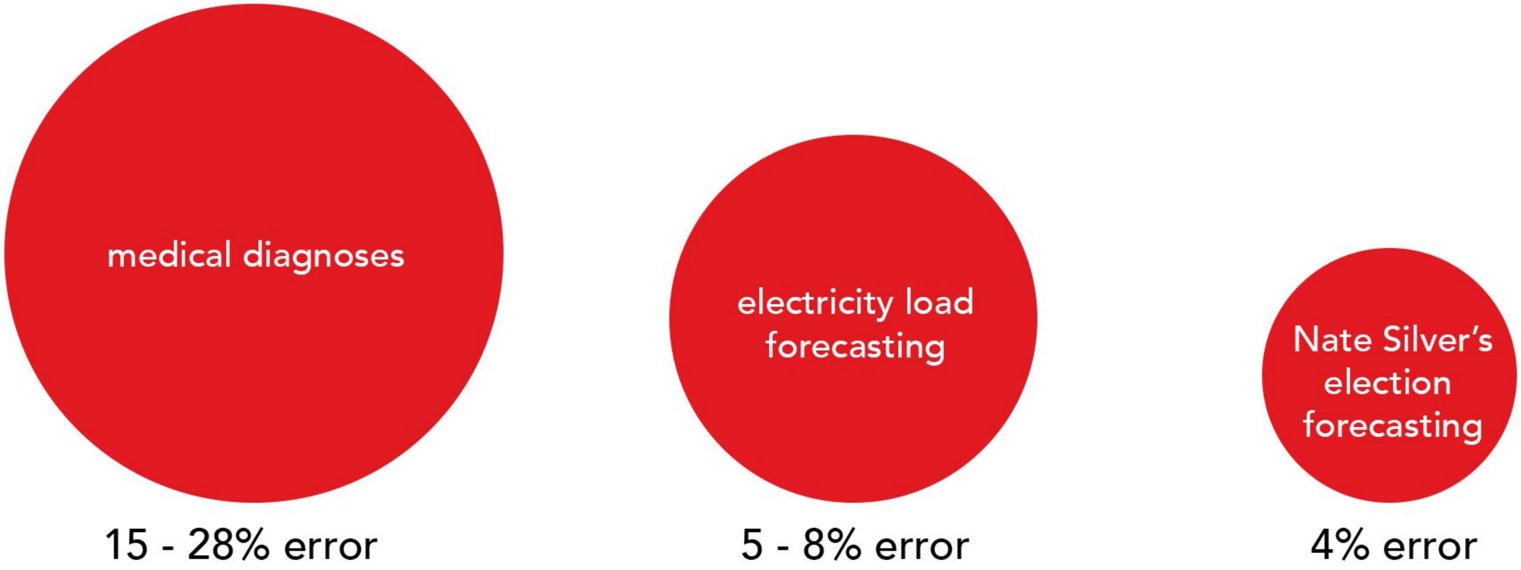


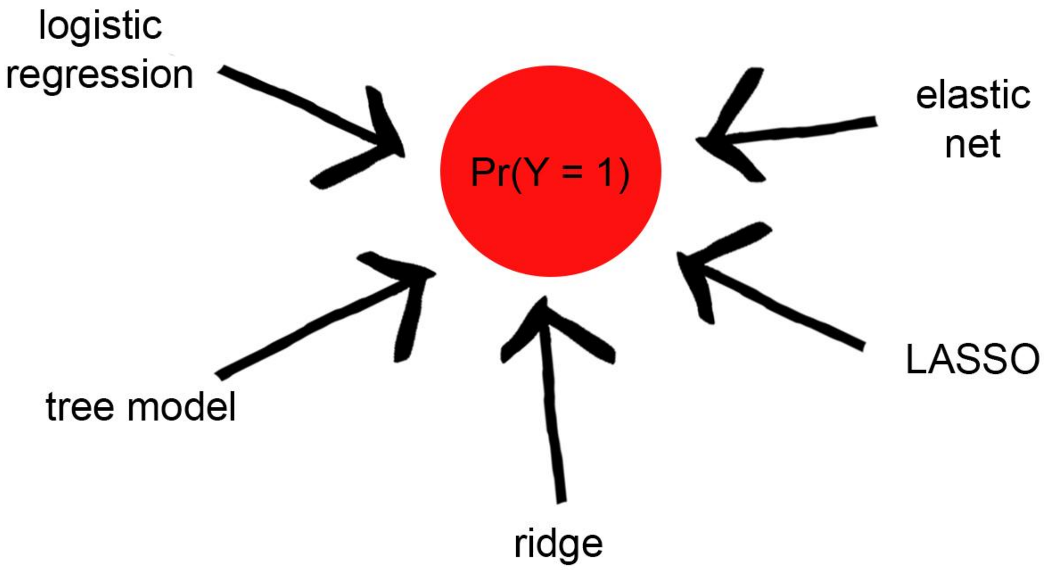
Image “Facebook gif” from Hackspot.blogspot.in

Room for Improvement in Diagnostic Accuracy



Berner et al., 2008, The American Journal of Medicine.
Gordon et al., 2009, Arch Intern Med.

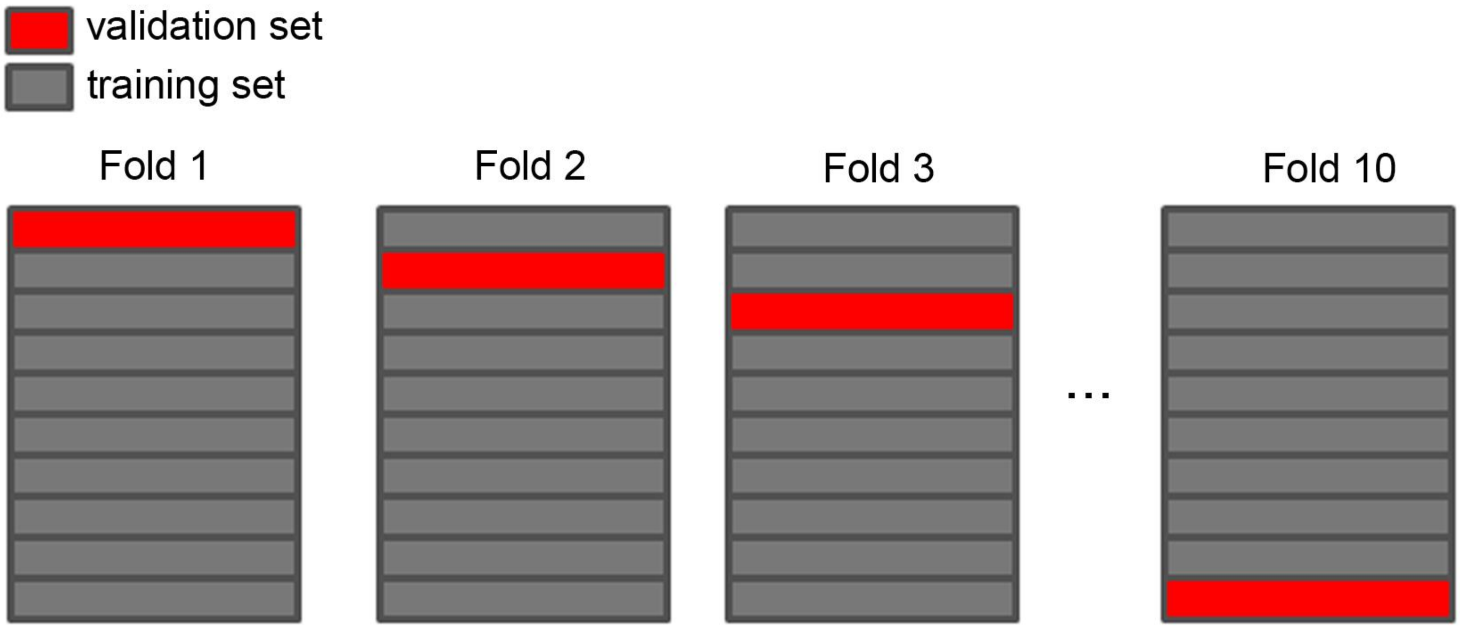
The Basics of Ensemble Machine Learning



Central question of ensemble learning: How much should each algorithm contribute to the prediction?

Dietterich, Multiple Classifier Systems: First International Workshop, 2000

Step 1: Determining Performance of the Algorithms



$$\text{CV MSE}_j = \frac{\sum_{i=1}^{n}(Y_i - D_{j,i})^2}{n}$$

Image “cv.png” from “Cross Validation and the Bias-Variance Trade-off (for Dummies)” from codesachin.wordpress.com
Screenshot of equation from Rose, 2013, American Journal of Epidemiology

Step 2: Optimizing the Aggregated Performance

Predicted values of each algorithm are used as the inputs in a regression to predict the actual outcomes, optimizing for minimum MSE. Coefficients become the weights on the individual algorithms.

$$\Pr(Y=1 \mid D) = \alpha_{lasso} D_{lasso} + \alpha_{ridge} D_{ridge} + \ldots + \alpha_{logistic} D_{logistic}$$

predicted values from the algorithm

estimated weight for the algorithm

actual outcomes

Step 2: Optimizing the Aggregated Performance

Resultant equation produces predictions of the Super Learner from the weighted aggregate of predictions from each individual algorithm.

$$\Pr(Y=1 \mid D) = 0.46D_{\text{lasso}} + 0.13D_{\text{ridge}} + \ldots + 0.27D_{\text{logistic}}$$

predicted values from the algorithm

weight for the algorithm

prediction of the Super Learner

Regression Methods

Logistic Regression

- Similar to standard regression but dependent variable is binary
- Coefficients represent effect on log-odds of event occurrence

Ridge Regression

- Reduces variance of model
- Shrinks/Penalizes coefficients

LASSO (least absolute shrinkage and selection operator)

- Can shrink variables to zero → variable selection

Elastic Net

- Compromise between Ridge and LASSO
- Variable selection + correlated predictors

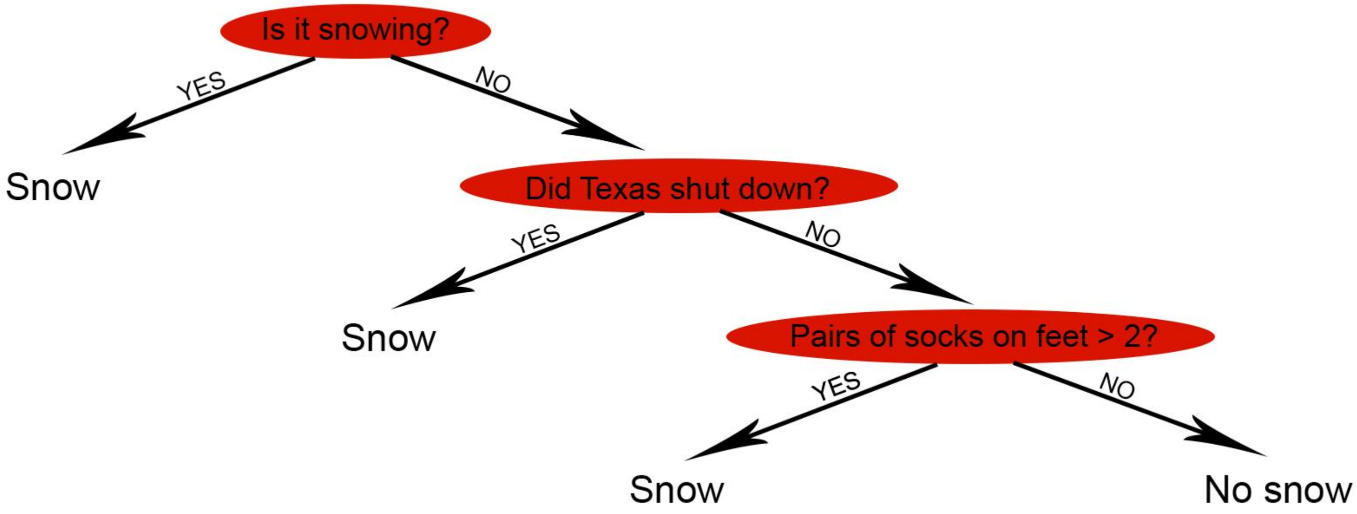
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

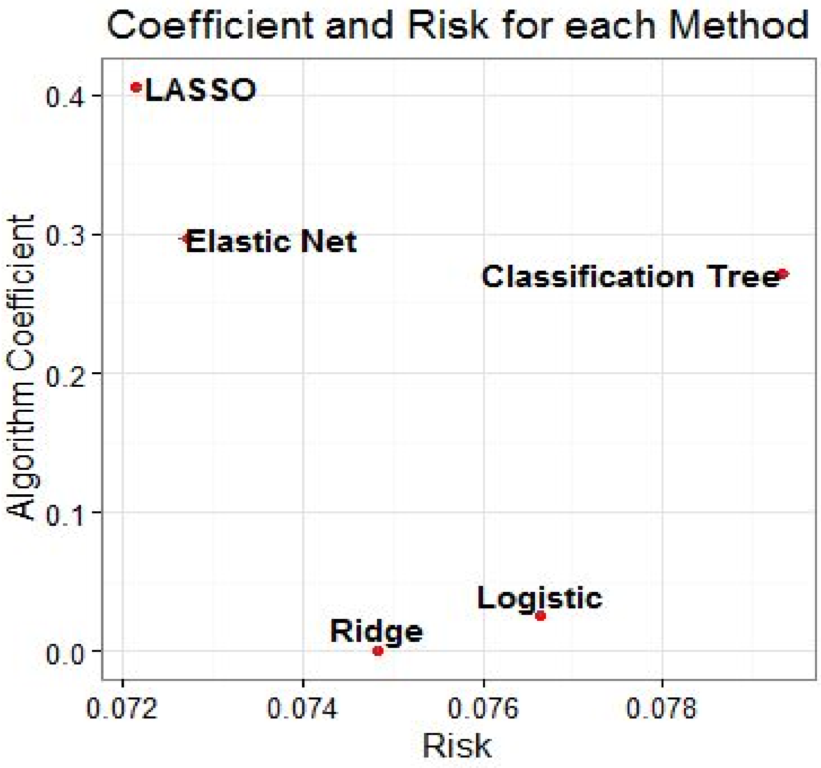
Tree-based Models

Classification Tree

- Each ‘node’ is a regressor
- Each ‘branch’ is a Y/N response



Super Learner Results vs. Included Algorithms



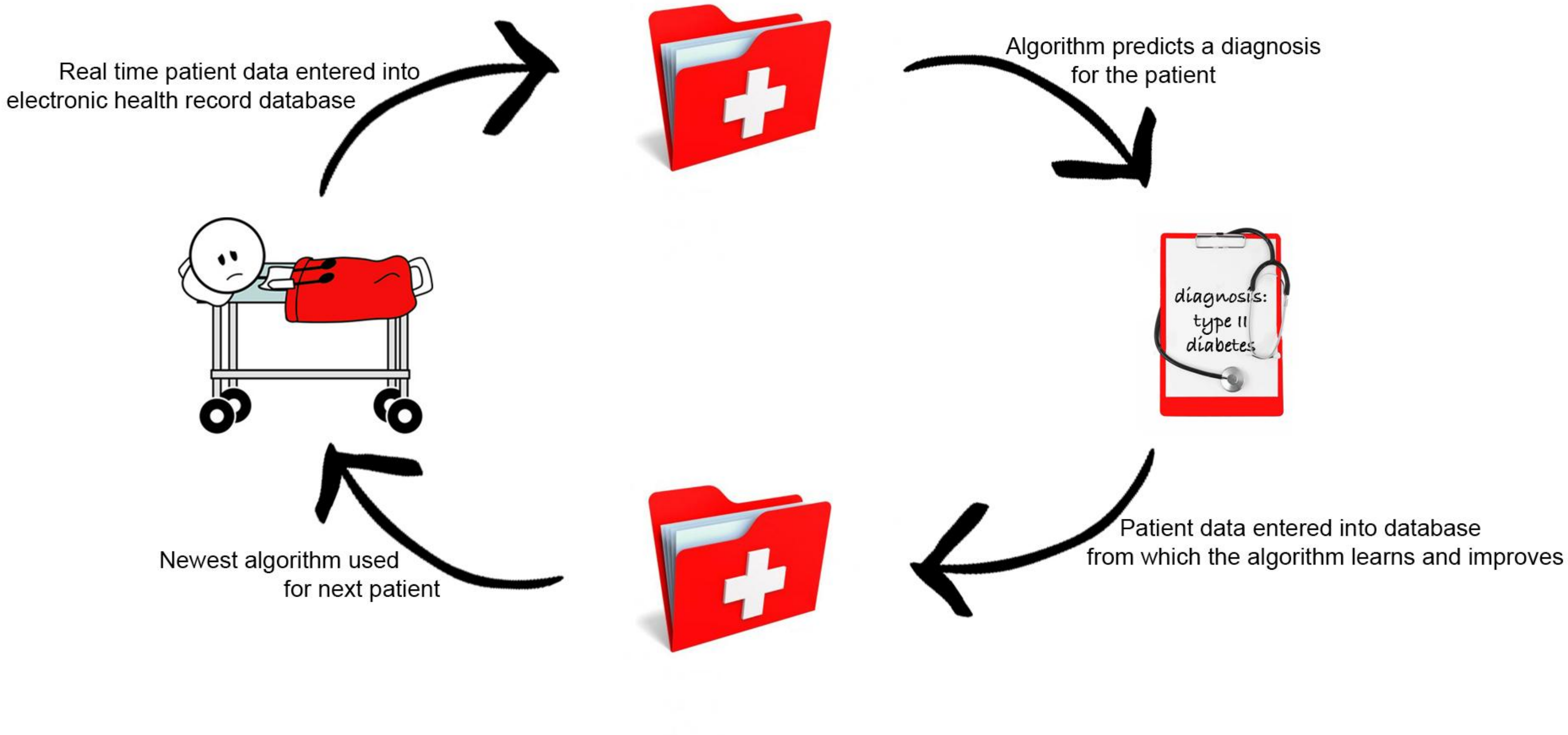
Algorithm	MSE	RE
Logistic	0.0766	1.141
Ridge	0.0748	1.114
LASSO	0.0721	1.074
Elastic Net	0.0727	1.083
Tree Model	0.0793	1.181
Super Learner	**0.0672**	**1**

Limitations

- Full EHR dataset
  - More predictors
  - Actual diagnosis
  - Representative
- Include additional algorithms
- Separate validation set

Bayesian generalized linear model
Classification and Regression Training
Leekasso
Neural nets
Ridge regression
Forward stepwise regression
Generalized additive models
Lasso and elastic net generalized linear models
Local regression
Non-Negative Least Squares (NNLS)
Recursive partitioning and regression trees
Regressive partitioning and regression trees with pruning
Gradient boosting method
Bagging classification trees
Logistic regression
Multivariate adaptive polyspline regression
Stepwise regression by Akaike information criterion (AIC)
Control forest
Mean algorithm
Random forest
Stepwise regression
Stepwise regression with interactions
Generalized linear model
Generalized linear model with interactions
Support vector machine
K-nearest neighbors algorithm

Ultimate Goal: Machine Learning Diagnostics in Real Time



Thank you for listening!
Any questions?

Stock image "medical bot" from dreamstime.com