Predicting Diabetes with Machine Learning

Group 15
Bryant Leal, Sunny Vidhani, Jazline Keli,
Qinwen Zhou, Hao He, Hyeon Gu Kim

Data Description

This dataset was obtained from Kaggle.com but originates from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). It is using 8 diagnostic measurements to predict whether a patient has diabetes.

768

females

21+

Yrs old

Pima Indian

heritage

Predictor and Outcome Variables

The number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, blood serum insulin, body mass index, diabetes pedigree, and age

	preg	plas	pres	skin	test	mass	pedi	age	class
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1

Model Selection

Decision Trees

KNN

Logistic Regression

- Stepwise
- Ridge Regression
- LASSO

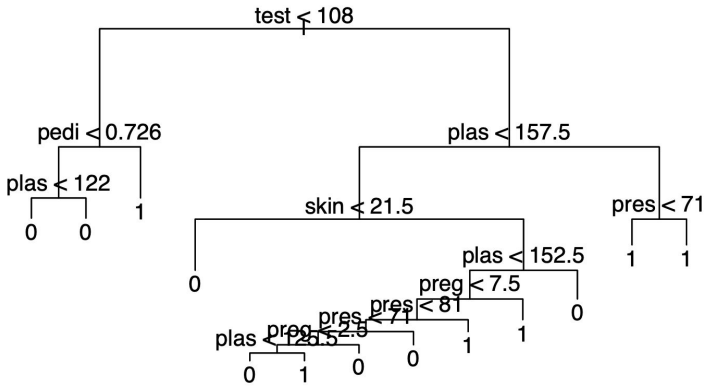
Decision Trees

Simple Tree

Random Forest

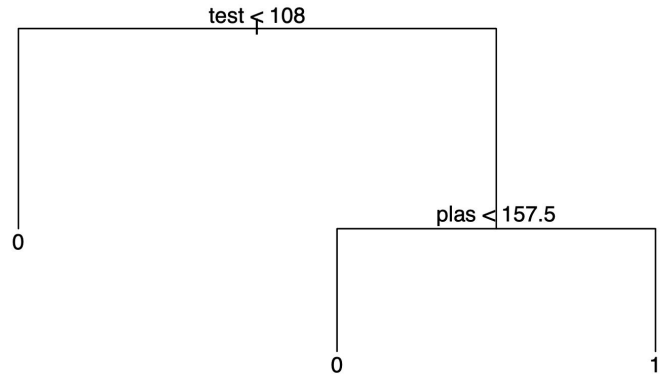
Simple Tree

The Big Tree: 13 nodes



Accuracy Rate: 75.19%

Pruned Tree: 3 nodes (optimal)



Accuracy Rate: 75.57%

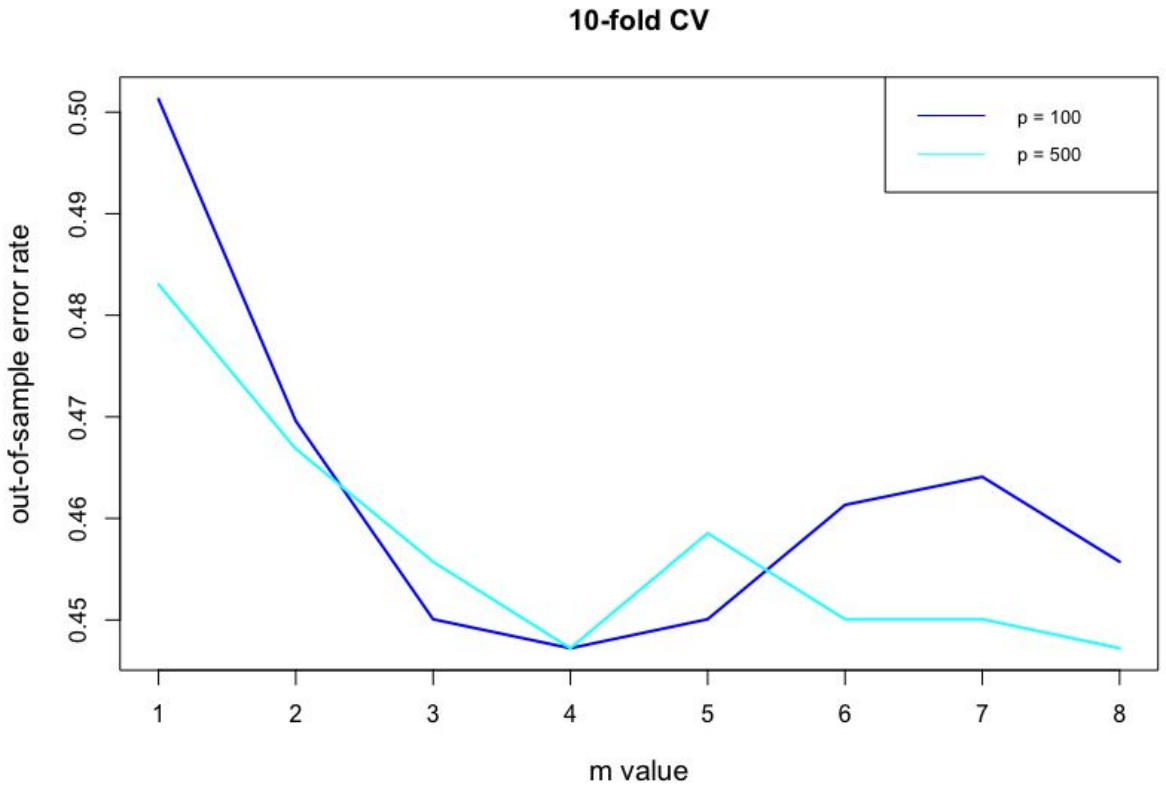
Random Forest

Optimal random forest model:

m = 4

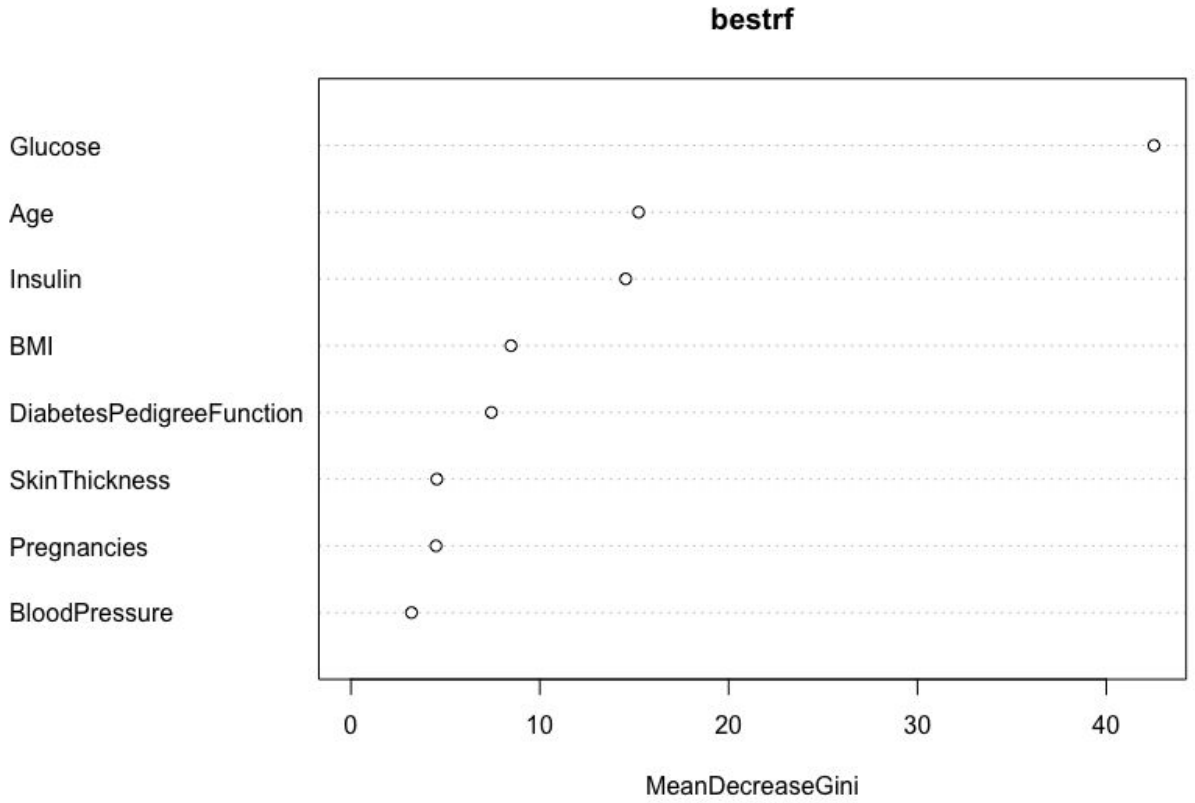
P = 100

Accuracy Rate: 77.55%

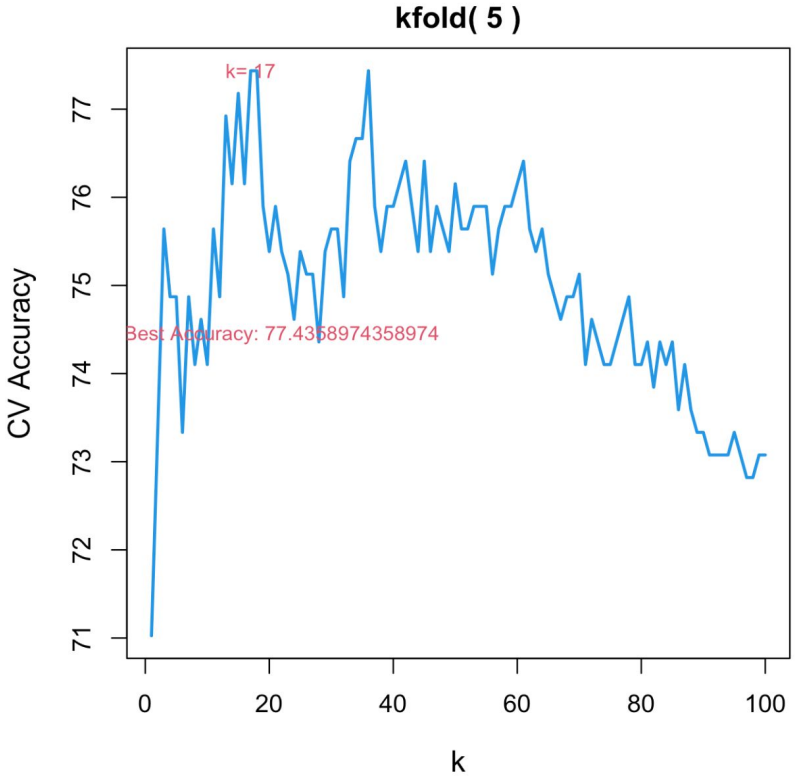


Random Forest

```
> importance(bestrf)
                          MeanDecreaseGini
Pregnancies                       4.502847
Glucose                          42.525873
BloodPressure                     3.207007
SkinThickness                     4.550779
Insulin                          14.548648
BMI                               8.474426
DiabetesPedigreeFunction          7.436104
Age                              15.231522
```



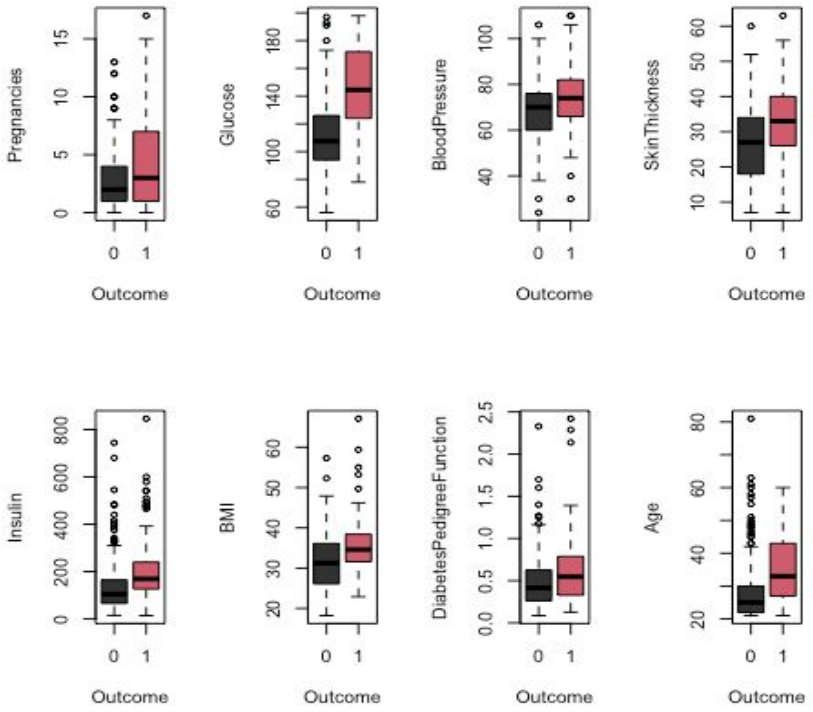
K-Nearest Neighbor



Logistic Regression

Cross Validation Accuracy: 77.60%

Black bar = outcomes for patients without* diabetes
Red bar = outcomes for patients with* diabetes



Stepwise Regression

class = β0 + β1(plas) + β2(mass) + β3(age) + ε

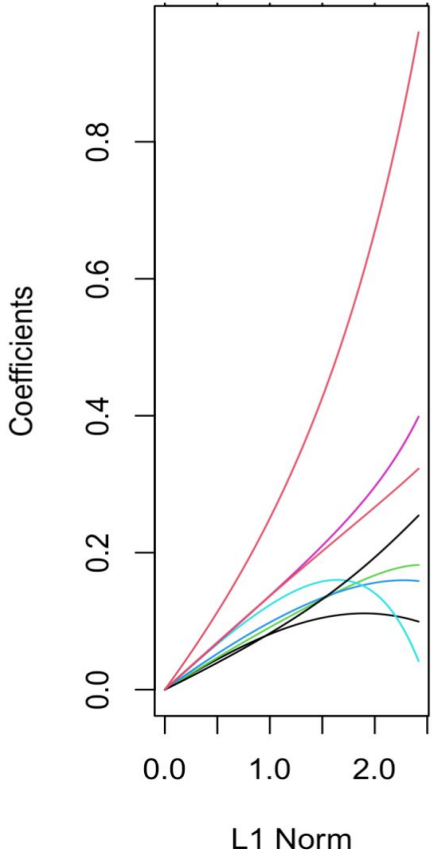
Accuracy = 76.27%

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6165  -0.6542  -0.3326   0.6944   2.3985

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9676     0.1714  -5.644 1.66e-08 ***
plas          1.2949     0.2017   6.420 1.36e-10 ***
mass          0.5979     0.1749   3.418 0.000631 ***
age           0.4145     0.1615   2.567 0.010246 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 353.67  on 273  degrees of freedom
Residual deviance: 243.45  on 270  degrees of freedom
AIC: 251.45
```

Ridge Variable Selection

Accuracy = 72.88%

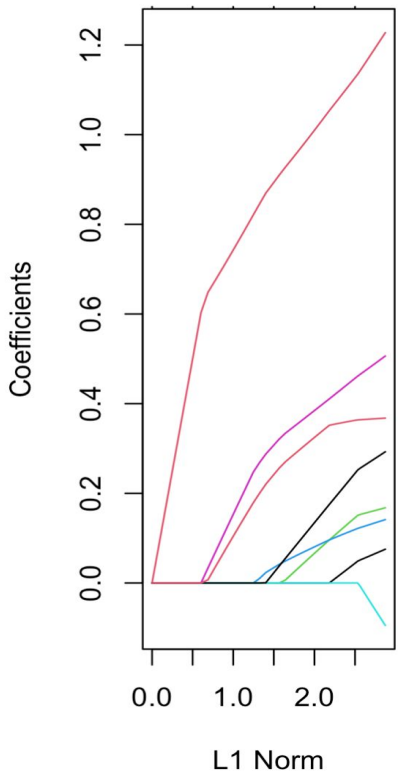
```
9 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -0.77240854
preg         0.09017058
plas         0.30087920
pres         0.10439730
skin         0.11076922
test         0.13790214
mass         0.15963130
pedi         0.09711465
age          0.15771673
```



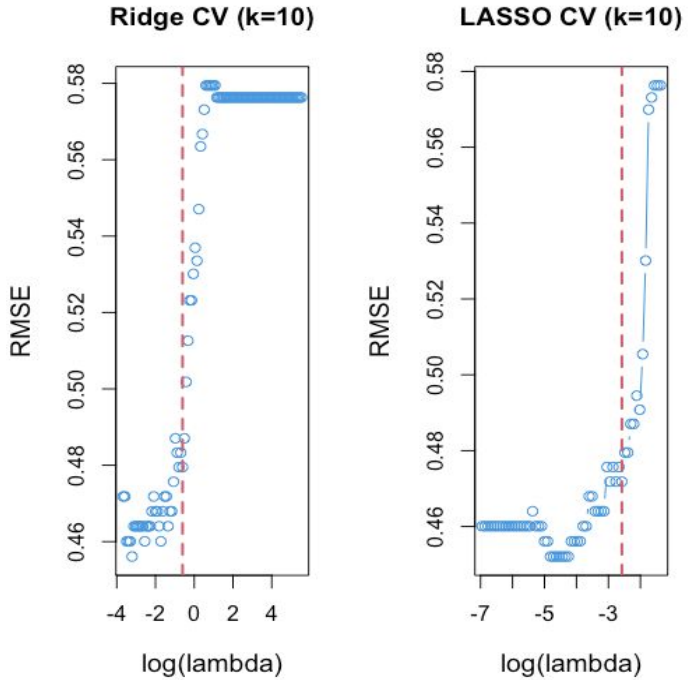
Lasso Variable Selection

Accuracy = 77.12%

```
9 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -0.8118152
preg         .
plas         0.7402151
pres         .
skin         .
test         .
mass         0.1500199
pedi         .
age          0.1017843
[1] 0.7711864
```



Lambda Graph



Conclusion

lasso_predict	test_y 0	1
0	75	23
1	4	16

glm_predict	test_y 0	1
0	69	15
1	10	24

Lasso

Out-of-Sample: 77.12%

Sensitivity: 41.03%
Specificity: 94.94%

Limitation of the Data
33.16% Positive
66% Negative

All Variables Regression

Out-of-Sample: 78.81%

Sensitivity: 61.54%
Specificity: 87.34%