Leveraging Machine Learning in R for Predicting Outcomes: A Case Study

Jagadish Katam

Introduction

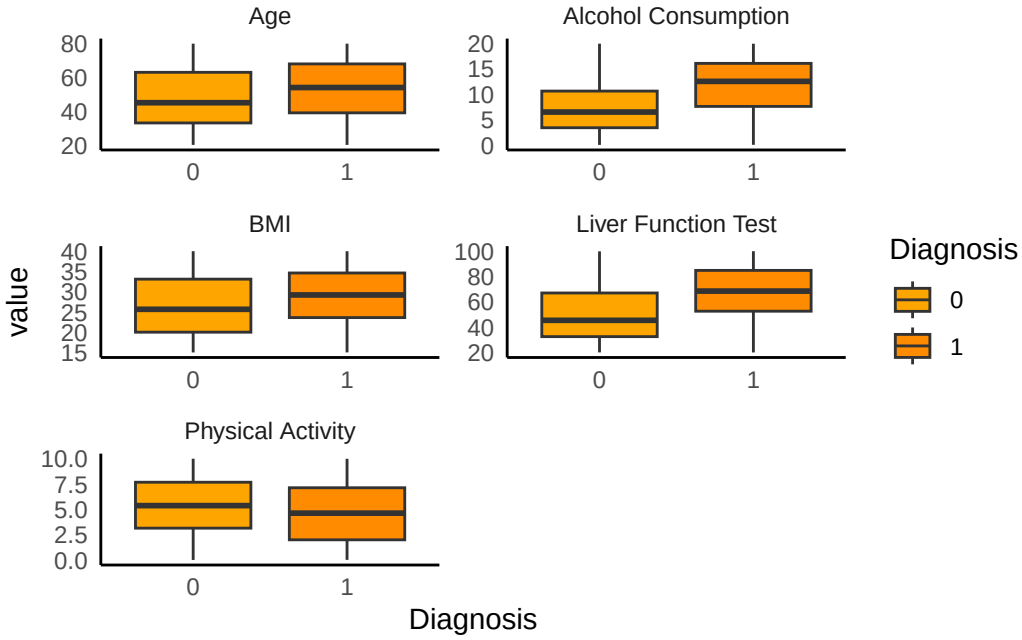
Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Data Science (DS) have been frequently interchangeably used. ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed.

Data

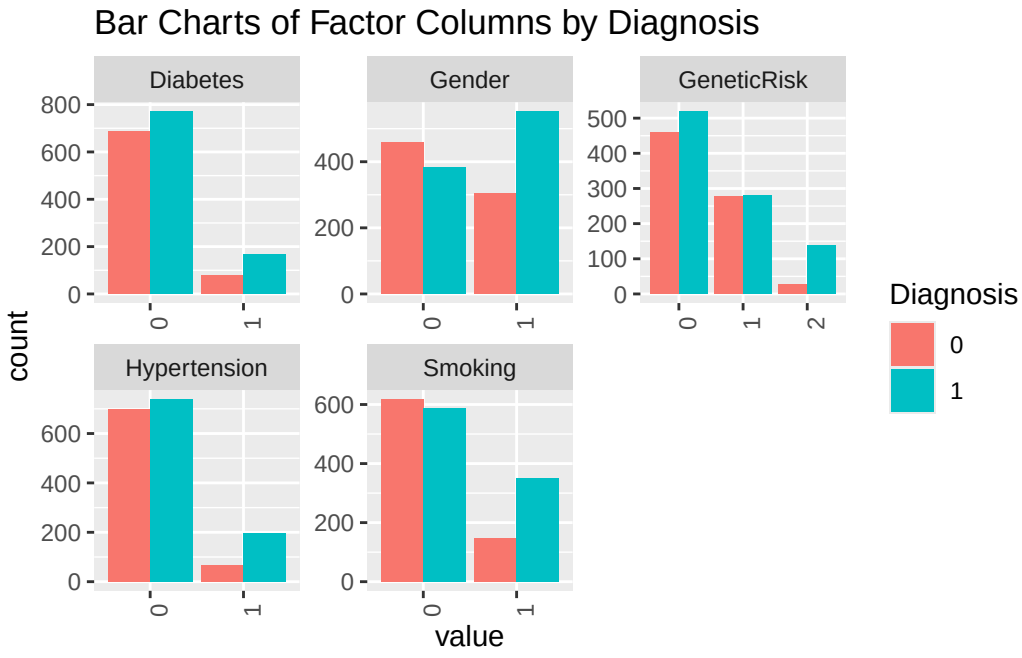
	Age	Gender	BMI	AlcoholConsumption	Smoking	GeneticRisk	PhysicalActivity
1	58	0	35.85758	17.272828	0	1	0.6589402
2	71	1	30.73247	2.201266	0	1	1.6705567
3	48	0	19.97141	18.500944	0	0	9.9283083
4	34	1	16.61542	12.632870	0	0	5.6301294
5	62	1	16.06583	1.087815	0	1	3.5662180
6	27	1	24.28521	12.885134	0	2	2.8820269

	Diabetes	Hypertension	LiverFunctionTest	Diagnosis
1	0	0	42.73424	1
2	1	0	67.30982	1
3	0	0	63.73896	0
4	0	0	64.55587	1
5	1	0	77.86869	1
6	0	0	50.53506	1

Box Plot



Bar Chart



Model Summary

```
Call:
stats::glm(formula = Diagnosis ~ Age + Gender + BMI + AlcoholConsumption +
    Smoking + GeneticRisk + Hypertension + LiverFunctionTest +
    PhysicalActivity + Diabetes, family = binomial, data = trainData)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -10.887044   0.720853 -15.103  < 2e-16 ***
Age                  0.037756   0.004956   7.619 2.56e-14 ***
Gender1              1.391007   0.173232   8.030 9.77e-16 ***
BMI                  0.076820   0.011688   6.572 4.95e-11 ***
AlcoholConsumption   0.247559   0.017858  13.863  < 2e-16 ***
Smoking1             1.760336   0.196054   8.979  < 2e-16 ***
GeneticRisk1        -0.140282   0.176799  -0.793    0.428
GeneticRisk2         2.573363   0.334900   7.684 1.54e-14 ***
Hypertension1        1.545425   0.253835   6.088 1.14e-09 ***
LiverFunctionTest    0.061146   0.004434  13.789  < 2e-16 ***
PhysicalActivity    -0.126914   0.029272  -4.336 1.45e-05 ***
Diabetes1            1.103333   0.260131   4.241 2.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1697.72  on 1234  degrees of freedom
Residual deviance:  934.08  on 1223  degrees of freedom
AIC: 958.08

Number of Fisher Scoring iterations: 6
```

Model Accuracy

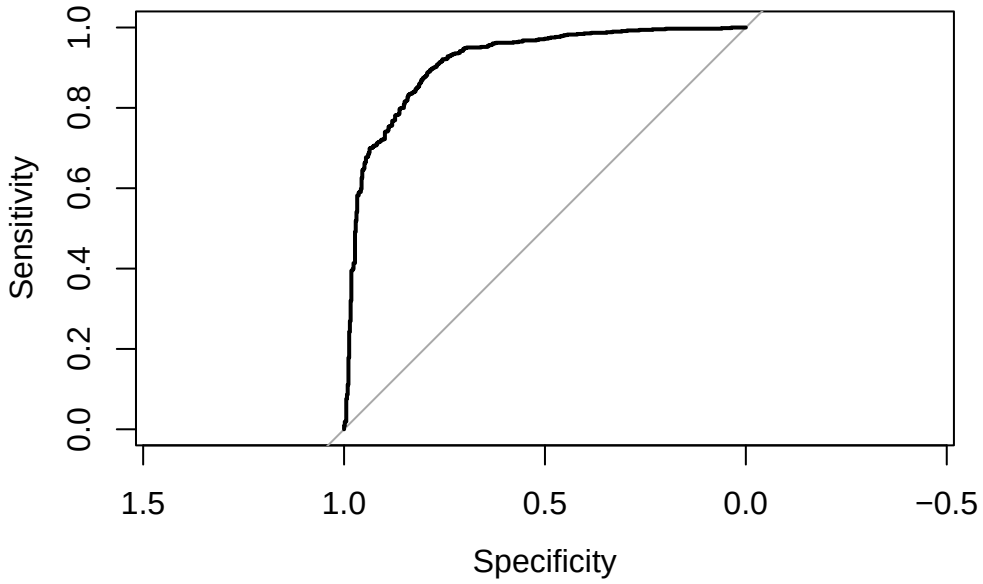
```
[1] 0.8516129
```

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 181  37
         1  32 215

               Accuracy : 0.8516
                 95% CI : (0.816, 0.8827)
    No Information Rate : 0.5419
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7017

 Mcnemar's Test P-Value : 0.6301

            Sensitivity : 0.8532
            Specificity : 0.8498
         Pos Pred Value : 0.8704
         Neg Pred Value : 0.8303
              Precision : 0.8704
                 Recall : 0.8532
                     F1 : 0.8617
             Prevalence : 0.5419
         Detection Rate : 0.4624
   Detection Prevalence : 0.5312
      Balanced Accuracy : 0.8515

       'Positive' Class : 1
```

Receiver Operating Characteristic (ROC)

```
Call:
roc.default(response = trainData$Diagnosis, predictor = model$fitted.values,     plot = TRUE)

Data: model$fitted.values in 551 controls (trainData$Diagnosis 0) < 684 cases (trainData$Diag
Area under the curve: 0.9146
```



Predicted Probabilities