Predicting Diabetes in the Pima Indians: An Investigation into Classification Strategies

May 9, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

this is the aim

2 Data

The dataset used throughout this paper originates from the National Institute of Diabetes and Digestive and Kidney Diseases and was first used in a demonstration of the ADAP Learning Algorithm in 1988 [1].

3 Results & Discussion

All results are 10-fold stratified cross validation accuracy figures in percentage (%).

Numeric Data	ZeroR	1R	1NN	5NN	NB	MLP	SVM	MyNB
No feature selection	65.1042	70.8333	67.8385	74.4792	75.1302	75.3906	76.3021	75.2614
CFS	65.1042	70.8333	69.0104	74.4792	76.3021	75.7813	76.6927	76.0407

Nominal Data	DT unpruned	DT pruned	MyDT	Bagg	Boost	RF
No feature selection	75	75.3906	73.4484	74.8698	76.1719	73.1771
CFS	79.4271	79.4271	78.3869	78.5156	78.6458	78.9063

J48 unpruned tree

---

a = high
| c = high
| | e = high: yes (82.0/31.0)
| | e = low: no (50.0/21.0)
| c = low: no (29.0/4.0)
a = low
| c = high
| | b = high
| | | e = high
| | | | d = high: yes (7.0/3.0)
| | | | d = low: no (28.0/4.0)
| | | e = low: no (43.0/4.0)
| | b = low: no (48.0/2.0)
| c = low: no (66.0)
a = very high
| b = high
| | c = high: yes (103.0/16.0)
| | c = low
| | | e = high: yes (12.0/3.0)
| | | e = low: no (4.0/1.0)
| b = low: no (3.0/1.0)
a = medium
| e = high
| | c = high
| | | d = high: yes (37.0/10.0)
| | | d = low: no (80.0/33.0)
| | c = low: no (30.0/3.0)
| e = low: no (146.0/17.0)

J48 pruned tree

---

a = high
| c = high
| | e = high: yes (82.0/31.0)
| | e = low: no (50.0/21.0)
| c = low: no (29.0/4.0)
a = low: no (192.0/14.0)
a = very high: yes (122.0/24.0)
a = medium
| e = high
| | c = high
| | | d = high: yes (37.0/10.0)

| | | d = low: no (80.0/33.0)
| | c = low: no (30.0/3.0)
| e = low: no (146.0/17.0)

3.1 Feature Selection

3.2 Comparison of Classifiers

4 Conclusion

conclusion

5 Reflection

References

[1] Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care 10* (11 1988).