ASSIGNMENT 1
ML TOOLS & TECHNIQUES

Srijan R. Shetty
(11727)

Salient Points

- ➔ The assignment was performed in **R, v3.1.2 - "Pumpkin Helmet".**
- ➔ The package **rpart** was used for the construction of Decision Trees.
- ➔ **Entropy** and **Gini** were used as entropy functions.

Handling Absurd Values

- ➔ The number of data vectors being only 768, leaving out vectors with absurd values was not a viable alternative.
- ➔ Absurd values were replaced with missing values - namely **'NA' in R** - and were handled by creating a **maximum of 5 surrogates** where possible.
- ➔ The following attribute values were **considered absurd** for the sake of this experiment by using consulting standard medical resources:
  - ◆ glucose concentration = 0
  - ◆ diastolic blood pressure <= 30
  - ◆ triceps skinfold thickness = 0
  - ◆ insulin = 0
  - ◆ bmi <= 10

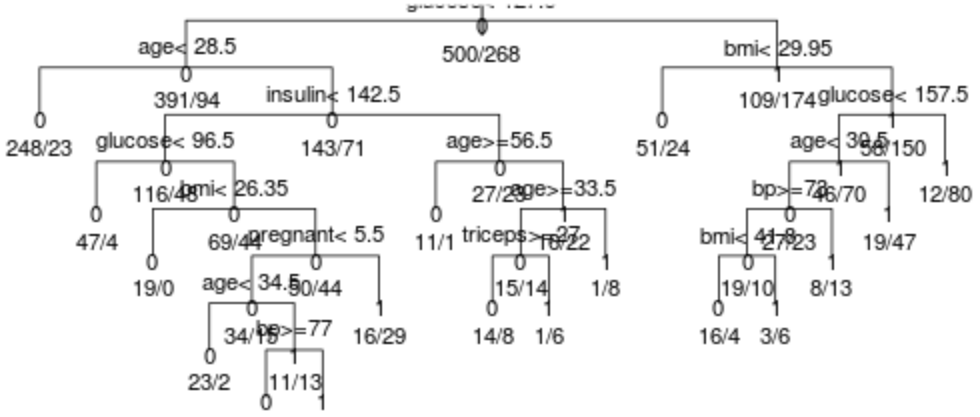
Results

- ➔ The five fold cross validation error can be reported using xerror which is the relative cross validation error with respect to the root node error.

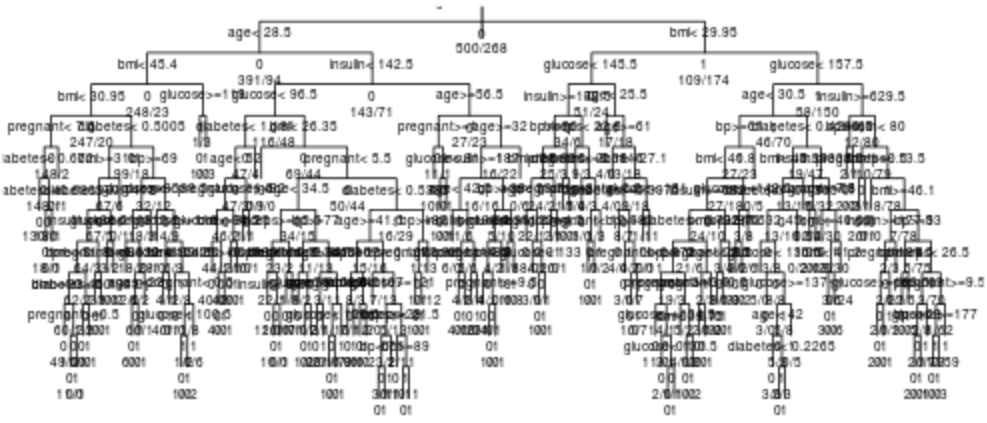
$$E = xerror \times root\ node\ error$$

	Threshold	Entropy	Gini
Threshold on vectors	minsplit=20, i.e., there must be atleast 20 vectors at a node to split	25.78%	24.21%
Threshold on impurity	cp=0.01, i.e., complexity (loss + penalty)	24.61%	23.34%
Grow tree and prune completely	**Complete Tree:** minsplit=1 **Prune**: cp=the value for which xerror is minimum	25.4%	23.82%

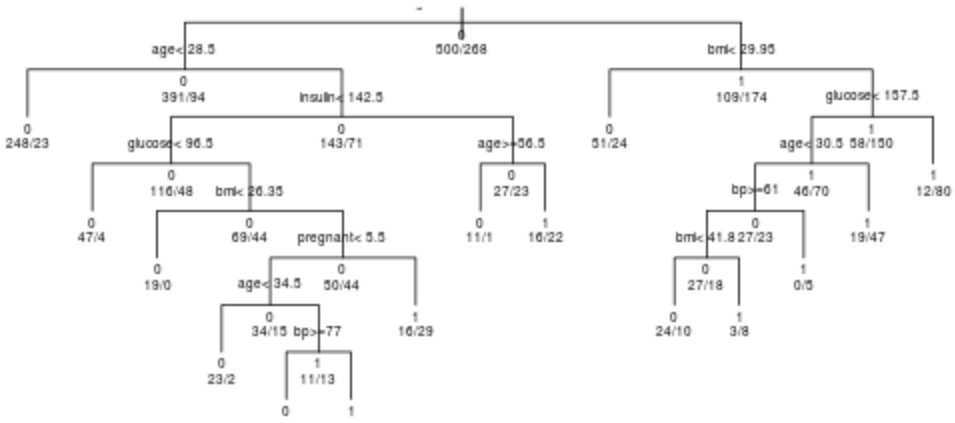
Threshold on Impurity - Gini



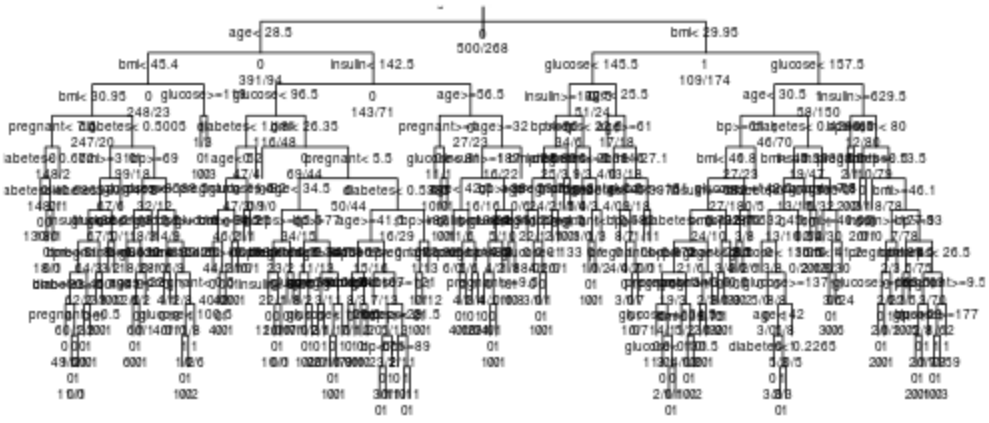
Complete Tree - Entropy



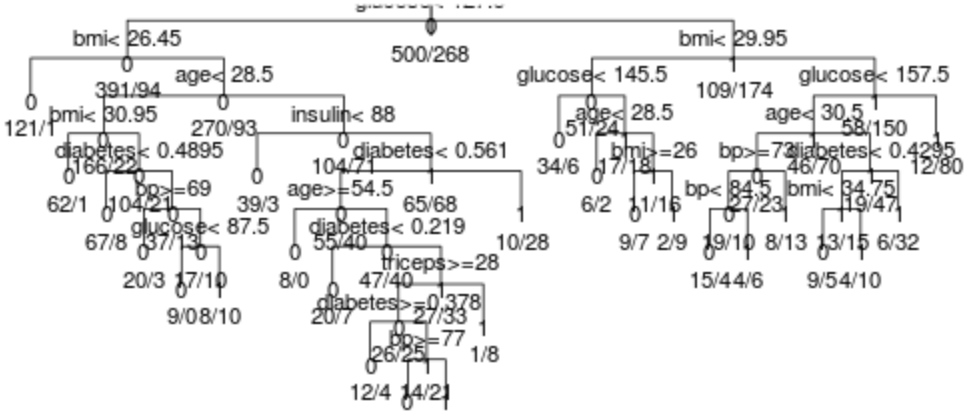
Pruned Tree - Entropy



Complete Tree - Gini



Threshold on vectors - Entropy



Threshold on vectors - Gini