1. Title: Poker Hand Dataset

2. Source Information

a) Creators:

Robert Cattral (cattral@gmail.com)

Franz Oppacher (oppacher@scs.carleton.ca)
Carleton University, Department of Computer Science
Intelligent Systems Research Unit
1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S5B6

c) Date of release: Jan 2007

3. Past Usage:
   1. R. Cattral, F. Oppacher, D. Deugo. Evolutionary Data Mining
      with Automatic Rule Generalization. Recent Advances in Computers,
      Computing and Communications, pp.296-300, WSEAS Press, 2002.
      - Note: This was a slightly different dataset that had more
              classes, and was considerably more difficult.

   - Predictive attribute: Poker Hand (labeled 'class')
   - Found to be a challenging dataset for classification algorithms
   - Relational learners have an advantage for some classes
   - The ability to learn high level constructs has an advantage

4. Relevant Information:
   Each record is an example of a hand consisting of five playing
   cards drawn from a standard deck of 52. Each card is described
   using two attributes (suit and rank), for a total of 10 predictive
   attributes. There is one Class attribute that describes the
   "Poker Hand". The order of cards is important, which is why there
   are 480 possible Royal Flush hands as compared to 4 (one for each
   suit – explained in more detail below).

5. Number of Instances: 25010 training, 1,000,000 testing

6. Number of Attributes: 10 predictive attributes, 1 goal attribute

7. Attribute Information:
   1) S1 "Suit of card #1"
      Ordinal (1-4) representing {Hearts, Spades, Diamonds, Clubs}

   2) C1 "Rank of card #1"
      Numerical (1-13) representing (Ace, 2, 3, ... , Queen, King)

   3) S2 "Suit of card #2"
      Ordinal (1-4) representing {Hearts, Spades, Diamonds, Clubs}

   4) C2 "Rank of card #2"
      Numerical (1-13) representing (Ace, 2, 3, ... , Queen, King)

   5) S3 "Suit of card #3"
      Ordinal (1-4) representing {Hearts, Spades, Diamonds, Clubs}

   6) C3 "Rank of card #3"
      Numerical (1-13) representing (Ace, 2, 3, ... , Queen, King)

   7) S4 "Suit of card #4"
      Ordinal (1-4) representing {Hearts, Spades, Diamonds, Clubs}

   8) C4 "Rank of card #4"
      Numerical (1-13) representing (Ace, 2, 3, ... , Queen, King)

   9) S5 "Suit of card #5"
      Ordinal (1-4) representing {Hearts, Spades, Diamonds, Clubs}

   10) C5 "Rank of card 5"

Numerical (1-13) representing (Ace, 2, 3, ... , Queen, King)

11) CLASS “Poker Hand”
Ordinal (0-9)

0: Nothing in hand; not a recognized poker hand
1: One pair; one pair of equal ranks within five cards
2: Two pairs; two pairs of equal ranks within five cards
3: Three of a kind; three equal ranks within five cards
4: Straight; five cards, sequentially ranked with no gaps
5: Flush; five cards with the same suit
6: Full house; pair + different rank three of a kind
7: Four of a kind; four equal ranks within five cards
8: Straight flush; straight + flush
9: Royal flush; {Ace, King, Queen, Jack, Ten} + flush

8. Missing Attribute Values: None

9. Class Distribution:

The first percentage in parenthesis is the representation
within the training set. The second is the probability in the full domain.

Training set:

0: Nothing in hand, 12493 instances (49.95202% / 50.117739%)
1: One pair, 10599 instances, (42.37905% / 42.256903%)
2: Two pairs, 1206 instances, (4.82207% / 4.753902%)
3: Three of a kind, 513 instances, (2.05118% / 2.112845%)
4: Straight, 93 instances, (0.37185% / 0.392465%)
5: Flush, 54 instances, (0.21591% / 0.19654%)
6: Full house, 36 instances, (0.14394% / 0.144058%)
7: Four of a kind, 6 instances, (0.02399% / 0.02401%)
8: Straight flush, 5 instances, (0.01999% / 0.001385%)
9: Royal flush, 5 instances, (0.01999% / 0.000154%)

The Straight flush and Royal flush hands are not as representative of
the true domain because they have been over-sampled. The Straight flush
is 14.43 times more likely to occur in the training set, while the
Royal flush is 129.82 times more likely.

Total of 25010 instances in a domain of 311,875,200.

Testing set:

The value inside parenthesis indicates the representation within the test
set as compared to the entire domain. 1.0 would be perfect representation,
while <1.0 are under-represented and >1.0 are over-represented.

0: Nothing in hand, 501209 instances,(1.000063)
1: One pair, 422498 instances,(0.999832)
2: Two pairs, 47622 instances, (1.001746)
3: Three of a kind, 21121 instances, (0.999647)
4: Straight, 3885 instances, (0.989897)
5: Flush, 1996 instances, (1.015569)
6: Full house, 1424 instances, (0.988491)
7: Four of a kind, 230 instances, (0.957934)
8: Straight flush, 12 instances, (0.866426)
9: Royal flush, 3 instances, (1.948052)

Total of one million instances in a domain of 311,875,200.

10. Statistics

Poker Hand	# of hands	Probability	# of combinations
Royal Flush	4	0.00000154	480
Straight Flush	36	0.00001385	4320