ASSIGNMENT – 2:

URL : http://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation

Abstract: *Data set containing values for 41 attributes (molecular descriptors) used to classify 1055 chemicals into 2 classes (ready and not ready biodegradable).*

Attribute Information:

41 molecular descriptors and *1 experimental class*:

1) SpMax_L: Leading eigenvalue from Laplace matrix

2) J_Dz(e): Balaban-like index from Barysz matrix weighted by Sanderson electronegativity

3) nHM: Number of heavy atoms

4) F01[N-N]: Frequency of N-N at topological distance 1

5) F04[C-N]: Frequency of C-N at topological distance 4

6) NssssC: Number of atoms of type ssssC

7) nCb-: Number of substituted benzene C(sp2)

8) C%: Percentage of C atoms

9) nCp: Number of terminal primary C(sp3)

10) nO: Number of oxygen atoms

11) F03[C-N]: Frequency of C-N at topological distance 3

12) SdssC: Sum of dssC E-states

13) HyWi_B(m): Hyper-Wiener-like index (log function) from Burden matrix weighted by mass

14) LOC: Lopping centric index

15) SM6_L: Spectral moment of order 6 from Laplace matrix

16) F03[C-O]: Frequency of C - O at topological distance 3

17) Me: Mean atomic Sanderson electronegativity (scaled on Carbon atom)

18) Mi: Mean first ionization potential (scaled on Carbon atom)

19) nN-N: Number of N hydrazines

20) nArNO2: Number of nitro groups (aromatic)

21) nCRX3: Number of CRX3

22) SpPosA_B(p): Normalized spectral positive sum from Burden matrix weighted by polarizability

23) nCIR: Number of circuits

24) B01[C-Br]: Presence/absence of C - Br at topological distance 1

25) B03[C-Cl]: Presence/absence of C - Cl at topological distance 3

26) N-073: Ar2NH / Ar3N / Ar2N-Al / R..N..R

27) SpMax_A: Leading eigenvalue from adjacency matrix (Lovasz-Pelikan index)

28) Psi_i_1d: Intrinsic state pseudoconnectivity index - type 1d

29) B04[C-Br]: Presence/absence of C - Br at topological distance 4

30) SdO: Sum of dO E-states

31) TI2_L: Second Mohar index from Laplace matrix

32) nCrt: Number of ring tertiary C(sp3)

33) C-026: R--CX--R

34) F02[C-N]: Frequency of C - N at topological distance 2

35) nHDon: Number of donor atoms for H-bonds (N and O)

36) SpMax_B(m): Leading eigenvalue from Burden matrix weighted by mass

37) Psi_i_A: Intrinsic state pseudoconnectivity index - type S average

38) nN: Number of Nitrogen atoms

39) SM6_B(m): Spectral moment of order 6 from Burden matrix weighted by mass

40) nArCOOR: Number of esters (aromatic)

41) nX: Number of halogen atoms

42) experimental class: ready biodegradable (RB) and not ready biodegradable (NRB) => Result(renamed)

The task was to predict if the chemicals are Biodegradable ready or not.

Parameter Range:

1. KNN :
   - K 1.0 11.0 11.0
   - K 41.0 61.0 21.0
   - K 1.0 41.0 21.0

2. J48 :
   - M 1.0 11.0 6.0
   - M 30.0 45.0 6.0
   - M 20.0 30.0 6.0

There are a lot of outliers and few have some unique values. The conclusion was made through Weka Visualization Section.

i) AUC Missing Data

ML Models	1%	5%	10%	20%
KNN	0.90	0.87***	0.81***	0.74***
J48	0.84**	0.84**	0.84*	0.83*
Naïve Bayes	0.89*	0.89	0.89	0.89
Neural network	0.90	0.85**	0.83**	0.80**

ii) AUC Noisy Data

ML Models	1%	5%	10%	20%
KNN	0.89	0.86	0.80**	0.75*
J48	0.84**	0.81**	0.78***	0.69***
Naïve Bayes	0.88*	0.85*	0.81*	0.72
Neural network	0.91	0.86	0.82	0.71**

Original AUC: Using Default Parameter.

ML Models	0%
KNN	0.92
J48	0.82
Naïve Bayes	0.89
Neural network	0.92

Observation and Conclusion:

Naïve Bayes and Neural Net run through default parameter

J48 – Works well with missing data and degrades the performance as nosie increase.

KNN – Linear drop in AUC with increasing noisy data and missing values

Neural – Linear decrease in AUC with increase in Noise and Missing data

Naïve Bayes – Works consistently well with missing data, though AUC decreases with increase in noise. With increasing missing values, Neural Net performance degrades for to low when compared with other models.

Naïve base seems to have consistent value with increase in missing data. Hence, can be a preferred method for dataset.