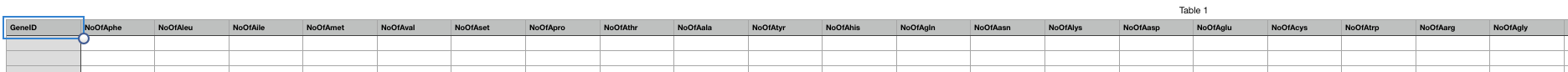
1. **Reprezentarea datelor ca frecventa aminoacizilor.**

Un mod de reprezentare al datelor pentru clasificare este de a parsa AND-ul intr-un sir de frecvente X[ai, ai+1, …, aN – 1, aN], ai = frecventa aminoacidului i.

Exemplu:



\*BnaC09g12820D 23 54 53 13 14 9 20 13 32 33 33 43 29 29 40 29 36 47 43 12 1 "gmp synthase"

Pentru aceasta reprezentare am rulat urmatorii algoritmi pentru a prezice proteina:

DecisionTreeClassifier

2-split

Cross val score: [0.48121737 0.50373832];

Score average : 0.49247784345764645

----------------------------------------------

RandomForestClassifier

2-split

Cross val score: [0.46440851 0.48531375];

Score average:0.4748611289861077

----------------------------------------------

ExtraTreesClassifier

2-split

Cross val score: [0.49713485 0.51922563];

Score average:0.5081802435506776

Daca am clasifica datele pe baza efectelor(efectul cel mai des intalnit pe tot setul de date) acuratetea ar fi peste 80%.

1. **Reprezentarea datelor ca sir de Aminoacizi in ordinea aparitiilor**

**Exemplu:**

BnaC09g12820D 20 16 12 19 18 13 10 12 15 9 18 13 2 14 14 8 2 16 15 3 15 10 13 11 9 6 3 4 19 11 1 6 3 11 3 16 3 9 7 19 15 2 18 6 19 17 19 3 3 13 19 17 11 14 12 10 14 12 11 9 16 15 18 15 14 16 17 9 2 19 2 18 10 3 17 9 2 10 18 2 4 18 3 2 3 15 16 18 3 3 16 3 19 19 3 9 15 3 10 10 16 3 16 12 16 2 3 7 18 17 2 12 17 12 18 6 3 3 2 15 5 20 3 1 15 1 13 11 19 18 19 11 12 19 19 16 2 19 13 19 18 7 11 6 2 5 13 16 8 12 16 17 3 14 15 1 12 10 3 2 2 16 14 3 16 3 2 2 12 15 15 5 1 7 1 2 19 19 15 2 18 11 19 12 12 18 8 11 12 18 17 1 13 15 11 9 5 15 14 12 2 1 2 18 16 8 7 4 19 12 11 3 12 10 3 17 4 9 18 11 1 17 7 2 13 3 17 14 3 11 7 12 1 19 12 18 16 1 7 12 12 3 12 3 15 1 18 9 18 2 18 17 8 17 3 1 11 13 19 19 19 16 3 9 15 9 3 6 9 1 15 19 18 18 2 13 13 5 15 8 9 7 15 13 3 3 17 5 15 3 9 12 1 15 19 19 12 12 19 10 15 3 5 13 7 10 14 3 10 7 9 10 2 19 9 11 9 18 9 3 13 10 2 13 20 17 17 17 15 8 3 9 19 15 11 3 12 11 1 14 12 12 8 9 2 15 2 19 7 9 1 9 13 1 17 14 10 3 3 1 12 10 6 10 10 2 7 2 3 16 14 12 2 19 18 14 3 9 13 20 2 14 14 11 15 1 18 15 7 10 20 8 9 14 18 2 2 2 7 17 12 7 18 17 13 20 1 14 15 19 17 18 12 19 13 8 16 2 3 18 10 14 5 10 18 14 2 13 17 10 1 20 12 2 10 3 11 11 7 3 11 3 14 11 17 14 15 3 11 2 17 17 12 11 17 4 2 9 18 2 9 10 10 15 9 18 2 18 3 5 10 18 4 12 17 12 18 20 9 20 7 20 14 10 14 14 7 18 17 18 17 18 10 16 3 16 13 5 2 13 12 14 10 14 18 3 17 15 13 15 19 18 17 18 7 16 17 12 19 4 16 19 2 12 17 13 2 10 11 19 3 1 19 13 18 15 15 11 18 2 10 5 16 6 10 18 19 11 2 2 18 11 20 17 15 11 3 8 2 13 2 17 3 8 3 15 9 3 3 12 19 17 19 2 2 5 2 12 15 13 5 14 8 14 19 12 13 11 16 12 16 19 20 11 9 18 16 2 9 10 15 16 6 11 5 stop "gmp synthase”

Fiecare CODON este mapat la un numar din intervalul [0, 20] pentru a reduce dimensiunea input-ului.

["TTT"] = 1 #"phe" ["TTC"] = 1 #"phe" ["TTA"] = 2 #"leu" ["TTG"] = 2 #"leu"  
["CTT"] = 2 #"leu" ["CTC"] = 2 #"leu" ["CTA"] = 2 #"leu" ["CTG"] = 2 #"leu"  
["TCT"] = 3 #"ser" ["TCC"] = 3 #"ser" ["TCA"] = 3 #"ser” ["TCG"] = 3 #"ser"  
["AGT"] = 3 #"ser" ["AGC"] = 3 #"ser" ["TAT"] = 4 #"tyr" ["TAC"] = 3 #"tyr"  
["TGT"] = 5 #"cys" ["TGC"] = 5 #"cys" ["TGG"] = 6 #"trp" ["CCT"] = 7 #"pro"  
["CCC"] = 7 #"pro" ["CCA"] = 7 #"pro" ["CCG"] = 7 #"pro" ["CAT"] = 8 #"his"  
["CAC"] = 8 #"his" ["CAA"] = 9 #"gln" ["CAG"] = 9 #"gln" ["CGT"] = 10 #"arg"  
["CGC"] = 10 #"arg" ["CGA"] = 10 #"arg ["AGA"] = 10 #"arg" ["AGG"] = 10 #"arg"  
["ATT"] = 11 #"ile"  
["ATC"] = 11 #"ile" ["ATA"] = 11 #"ile” ["ACT"] = 12 #"thr" ["ACC"] = 12 #"thr"  
["ACA"] = 12 #"thr" ["ACG"] = 12 #"thr" ["AAT"] = 13 #"asn" ["AAC"] = 13 #"asn"  
["AAA"] = 14 #"lys" ["AAG"] = 14 #"lys" ["GTT"] = 15 #"val" ["GTC"] = 15 #"val"  
["GTA"] = 15 #"val" ["GTG"] = 15 #"val" ["GCT"] = 16 #"ala" ["GCC"] = 16 #"ala"  
["GCA"] = 16 #"ala" ["GCG"] = 16 #"ala" ["GAT"] = 17 #"asp" ["GAC"] = 17 #"asp"  
["GAA"] = 18 #"glu" ["GAG"] = 18 #"glu" ["GGT"] = 19 #"gly" ["GGC"] = 19 #"gly"  
["GGA"] = 19 #"gly" ["GGG"] = 19 #"gly" ["ATG"] = 20 #"met" ["TAA"] = "stop" ["TAG"] = "stop" ["TGA"] = "stop"

\*”stop” indica finalul, iar imediat dupa e indicate proteina sintetizata de combinatia de aminoacizi.

Pe aceasta reprezentare s-au rulat urmatorii algoritmi(inclusive cei rulati la pasul anterior, rata de clasificare fiind aproximativ 10-15%) pentru a prezice proteina:

ZeroR :

Correctly Classified Instances 11.5686 %

Incorrectly Classified Instances 88.4314 %

Multinomial naive bayes for text data:

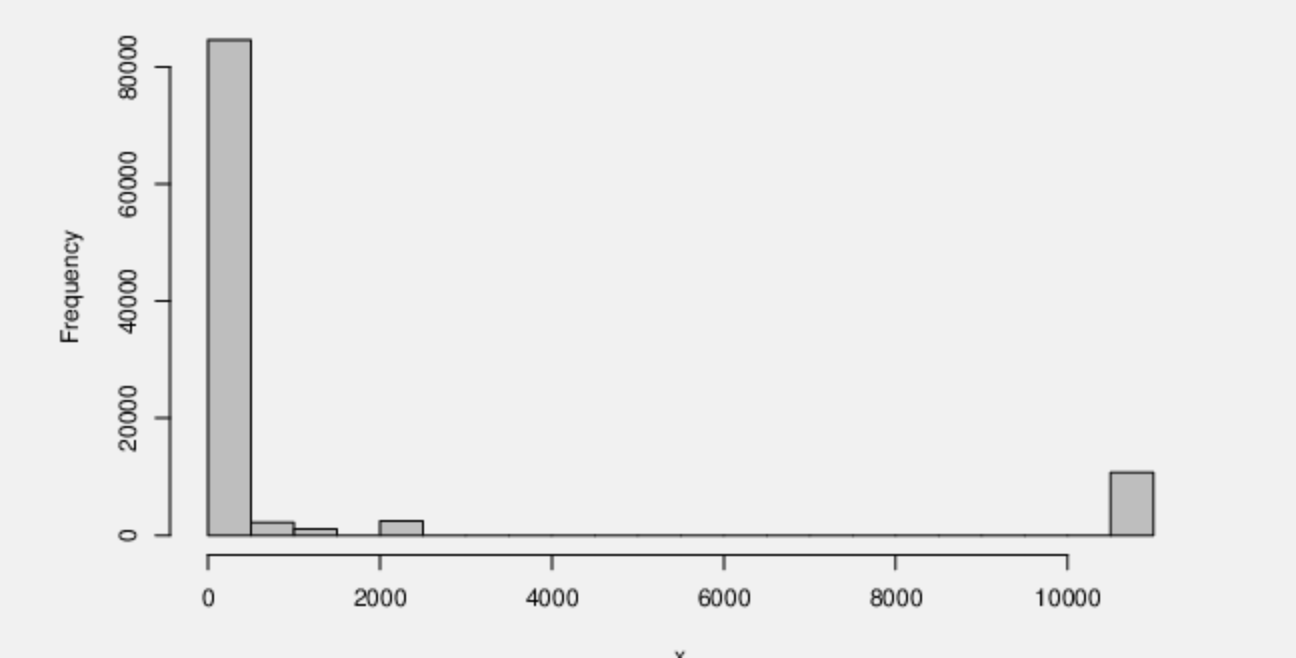
Correctly Classified Instances 10.5687 %

Incorrectly Classified Instances 89.4313 %

Se observa ca ambele reprezentari duc la o rata destul de mica, explicatia este una simpla:

Date insuficiente pentru ca antrena un model.

Distributia frecventelor proteinelor este urmatoarea



Se observa ca foarte multe proteine apar de putine ori, iar cam 20% apar 1 singura data.

Daca am clasifica datele pe baza efectelor(cel mai intalnit efect din tot setul de date) acuratetea creste la 25%.