

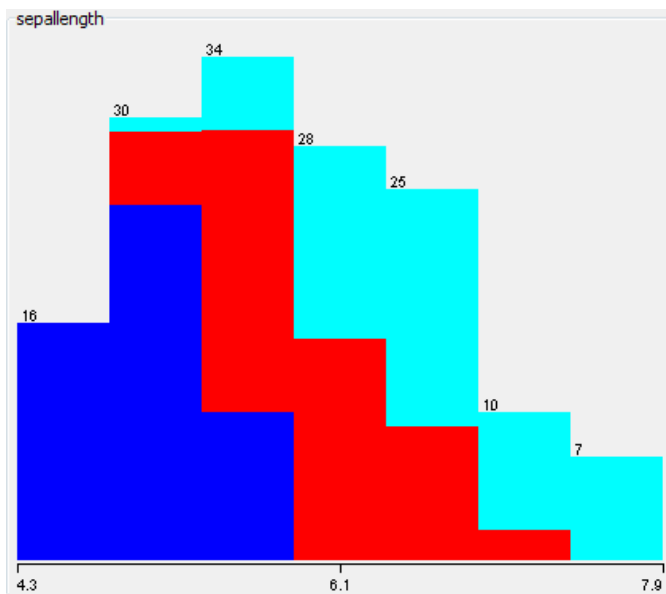
ΕΡΓΑΣΙΑ 02

Β ΜΕΡΟΣ

Το αρχείο iris περιέχει πληροφορίες για 150 φυτά τύπου iris και σύμφωνα με 4 παράγοντες κατατάσσονται τα φυτά σε 3 είδη. Η μεταβλητή στόχος είναι nominal και δέχεται ως ορίσματα τα iris-setosa, iris-versicolor και iris-virginica, που είναι τα είδη των φυτών. Οι παράγοντες που μελετάμε είναι οι εξής: sepallength, sepalwidth, petallength, petalwidth.

Με τη βοήθεια του λογισμικού weka μπορούμε να κάνουμε οπτική εξερεύνηση των δεδομένων. Στην ανάλυση που ακολουθεί το κόκκινο χρώμα σημαίνει ότι το φυτό ανήκει στην κατηγορία iris versicolor, το μπλε ότι ανήκει στην κατηγορία iris setosa και το θαλασσί ότι ανήκει στην κατηγορία iris virginica. Επίσης πρέπει να επισημάνουμε ότι επειδή τα attribute είναι όλα numerical τα παρακάτω ιστογράμματα δείχνουν την κατανομή του class attribute σαν συνάρτηση του κάθε attribute.

a) sepallength

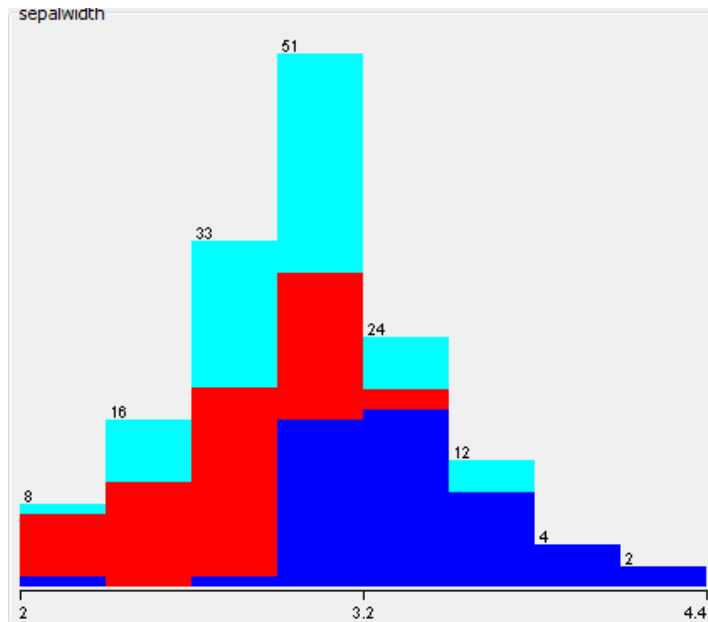


Στο παραπάνω σχήμα παρατηρούμε ότι 16 φυτά που έχουν μήκος σεφάλου 4,3-4,814 ανήκουν στην ομάδα iris setosa. 30 φυτά που έχουν μήκος σεφάλου 4,814-5,329 ανήκουν και στις 3 κατηγορίες. 7 φυτά που έχουν μήκος σεφάλου από 7,386- 7,9 ανήκουν στην κατηγορία iris virginica. Παρόμοια συμπεράσματα μπορούμε να εξάγουμε και για τις υπόλοιπες στήλες του γραφήματος.

Selected attribute		
Name: sepallength	Distinct: 35	Type: Numeric
Missing: 0 (0%)		Unique: 9 (6%)
Statistic	Value	
Minimum	4.3	
Maximum	7.9	
Mean	5.843	
StdDev	0.828	

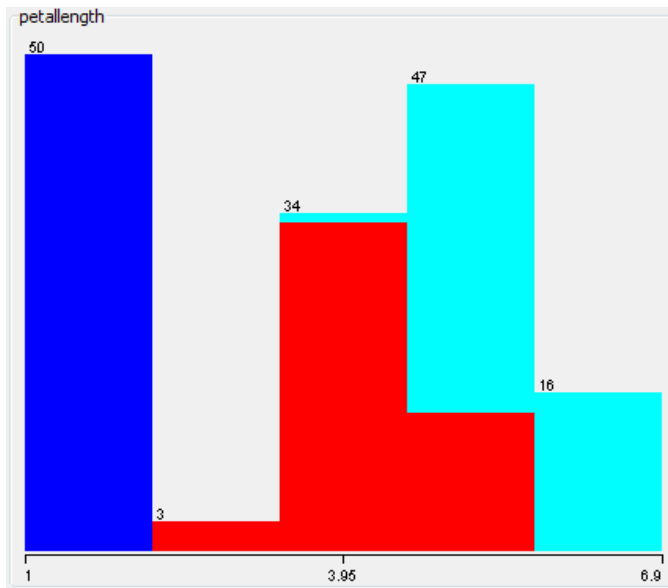
Το weka μας δίνει την δυνατότητα να δούμε για κάθε attribute το είδος της μεταβλητής (Numeric) ,αν υπάρχουν ελλειπούσες τιμές (missing) ,unique (στην περιπτωσή μας είναι 9) , πόσες distinct values υπάρχουν (35) καθώς και διάφορα στατιστικά στοιχεία όπως το μέγιστο, το ελάχιστο η μέση τιμή και η τυπική απόκλιση.

b) sepalwidth



Το attribute sepalwidth είναι και αυτό numeric, έχουμε 0 missing values ,23 distinct values και 5 unique values. Ομοίως μπορούμε σύμφωνα με το εύρος των σεφάλων να βγάλουμε συμπεράσματα για την κατηγορία που ανήκει κάθε φυτό. Επί παραδείγματι για τιμές από 3,8-4,4 έχουμε 6 φυτά να ανήκουν στην κατηγορία iris – setosa.

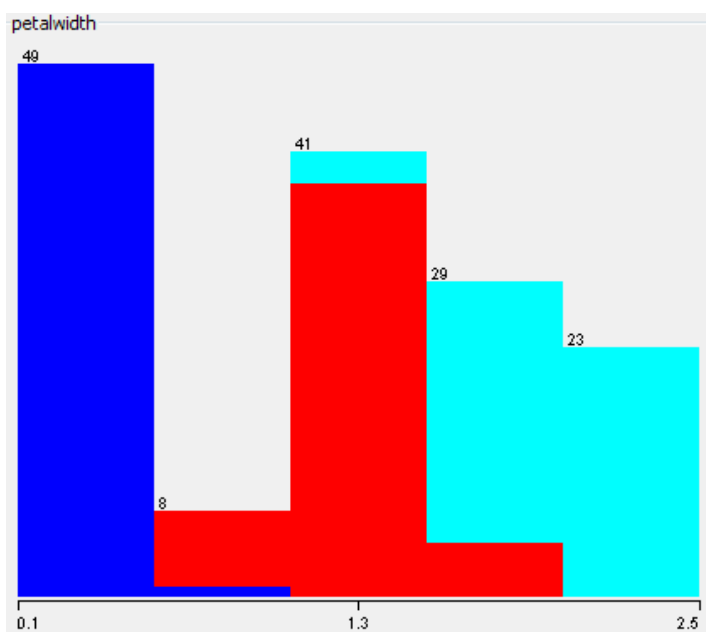
c) petallength



Όσον αφορά τη μεταβλητή μήκος πετάλου παρατηρούμε ότι το σύνολο των φυτών (50) που ανήκουν στην κατηγορία *iris-setosa* έχουν μήκος πετάλου από 1-2,18. Επίσης 16 φυτά με μήκος πετάλου από 5,72-6,9 ανήκουν στην κατηγορία *iris virginica* και 3 φυτά με μήκος πετάλου 2,18- 3,36 ανήκουν στην κατηγορία

To attribute *petalength* είναι numeric, έχει 0 missing values, 43 distinct values και 10 unique values.

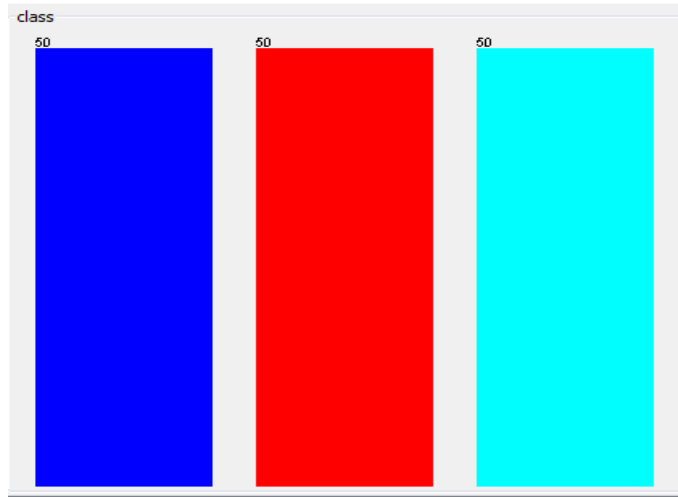
d) petalwidth



Για το εύρος του πετάλου παρατηρούμε πάλι ότι 49 φυτά που ανήκουν στην κατηγορία *iris setosa* έχουν εύρος 0,1-0,58 καθώς και ότι τα περισσότερα φυτά που ανήκουν στην κατηγορία *iris virginica* έχουν εύρος πετάλου από 1,54 έως 2,5.

Η μεταβλητή *petalwidth* είναι και αυτή numeric, έχει 0 missing values, 22 distinct values και 2 unique values.

e) Class attribute



Όπως βλέπουμε το σύνολο του πληθυσμού των φυτών είναι ίσα κατανεμημένο στις 3 ομάδες. 50 φυτά ανήκουν στο *iris setosa* 50 φυτά στο *iris versicolor* και 50 φυτά στο *iris virginica*.

Με την οπτική διερεύνηση των δεδομένων προκύπτουν διάφοροι κανόνες όπως

If *sepalwidth* > 6.781 then class= 'iris virginica'

If *sepalwidth* > 3.8 then class= 'iris setosa'

If *petalwidth* < 2.18 then class= 'iris setosa'

If *petalwidth* > 2.02 then class= 'iris virginica'