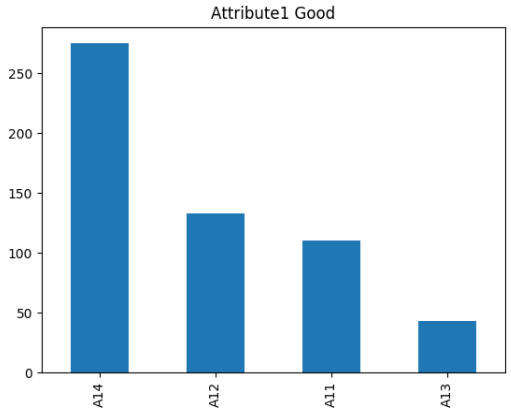
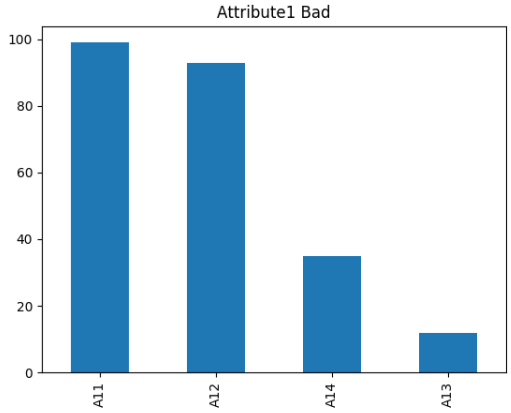
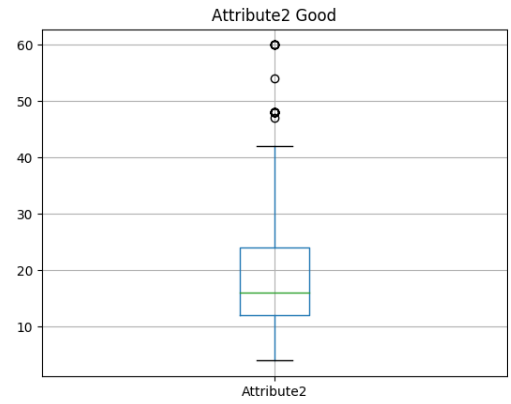
ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ
ΕΡΓΑΣΙΑ 2η

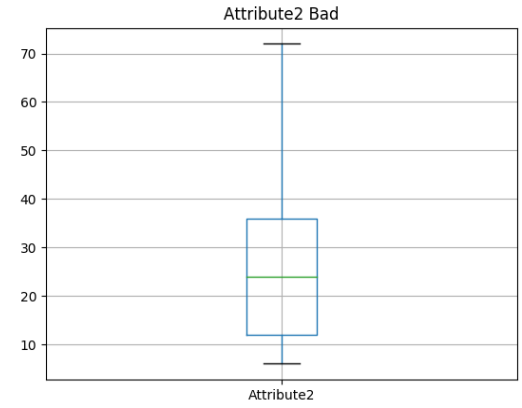
Συμμετέχοντες:
Βαγγέλης Τσιατούρας (1115201200185)
Νίκος Σοφράς (1115201200168)

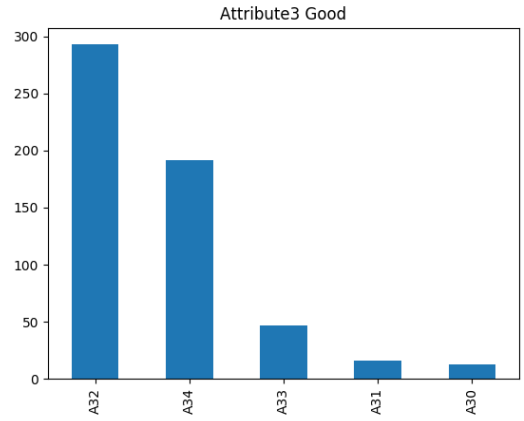
Οπτικοποίηση Δεδομένων

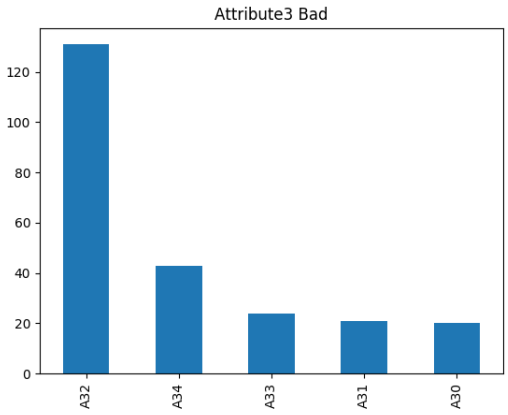


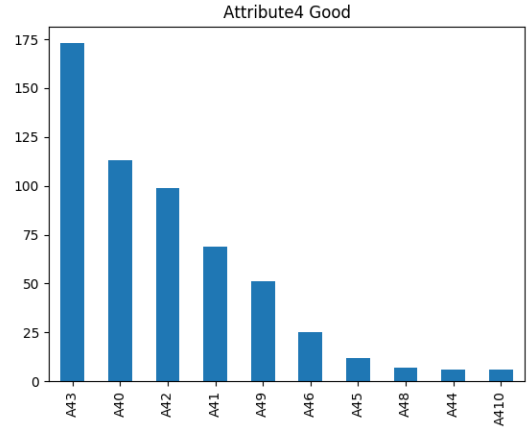


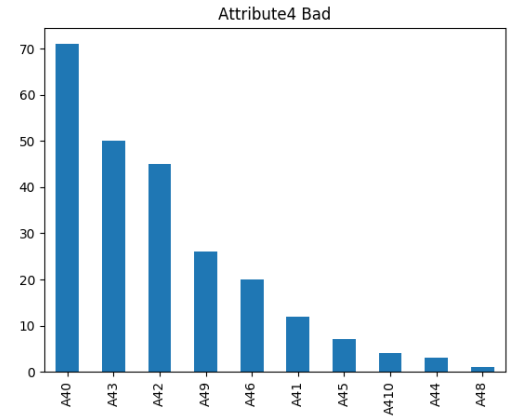


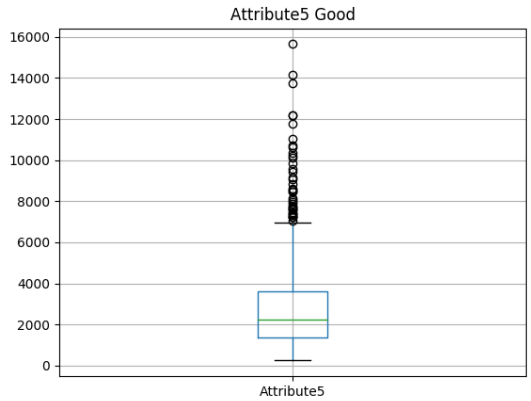


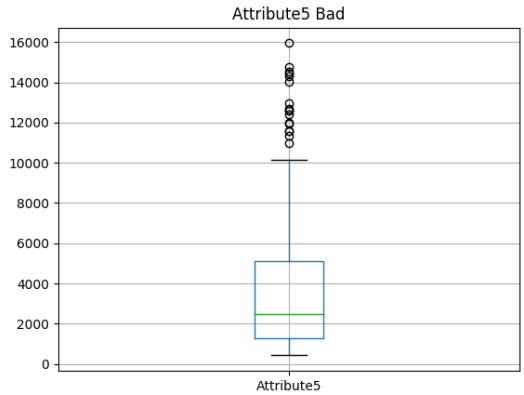

Attribute3 Good
300
250
200
150
100
50
0
A32
A34
A33
A31
A30

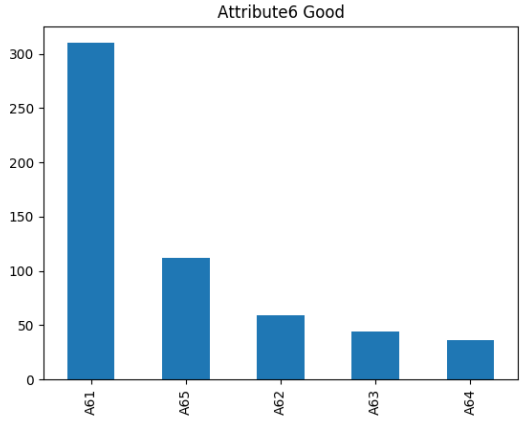

Attribute3 Bad
120
100
80
60
40
20
0
A32
A34
A33
A31
A30

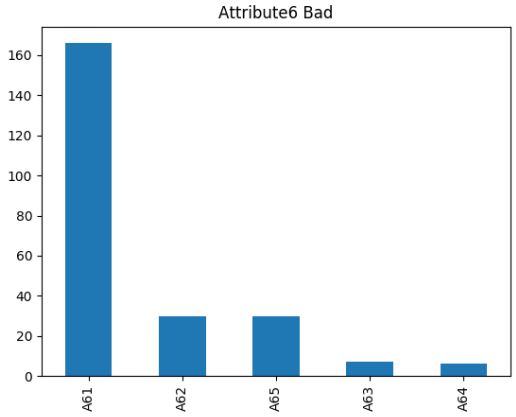

Attribute4 Good
175
150
125
100
75
50
25
0
A43
A40
A42
A41
A49
A46
A45
A48
A44
A410

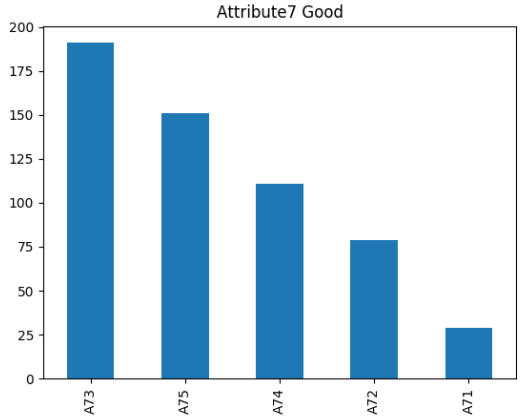

Attribute4 Bad
70
60
50
40
30
20
10
0
A40
A43
A42
A49
A46
A41
A45
A410
A44
A48

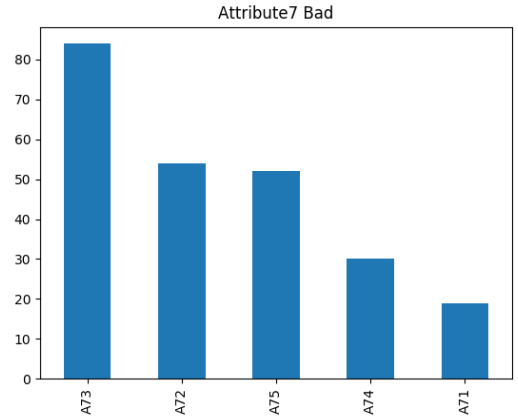

Attribute5 Good
16000
14000
12000
10000
8000
6000
4000
2000
0
Attribute5

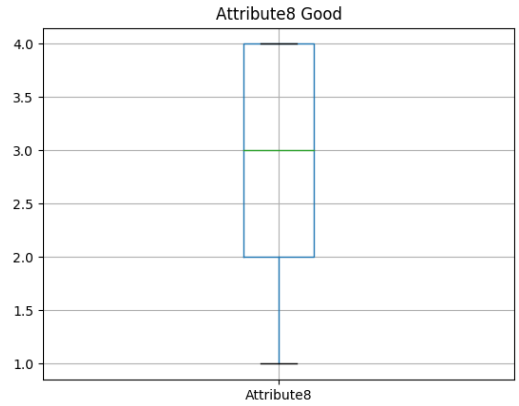

Attribute5 Bad
16000
14000
12000
10000
8000
6000
4000
2000
0
Attribute5

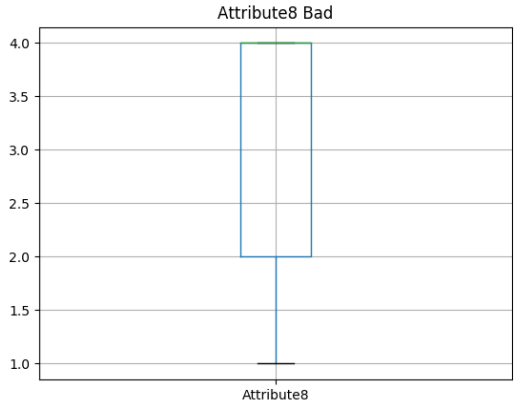

Attribute6 Good
300
250
200
150
100
50
0
A61
A65
A62
A63
A64

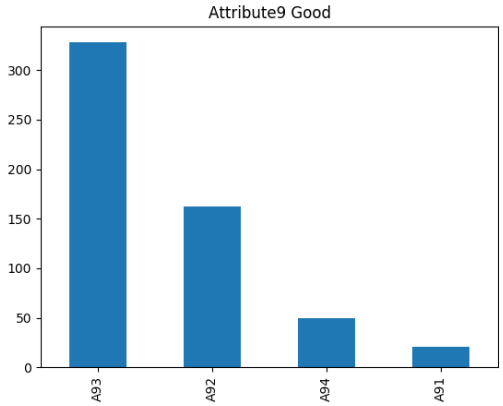

Attribute6 Bad
160
140
120
100
80
60
40
20
0
A61
A62
A65
A63
A64

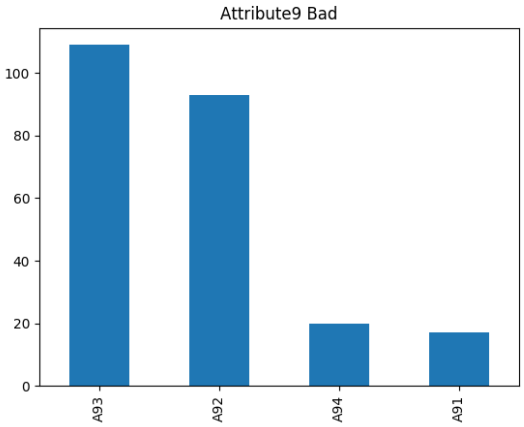

Attribute7 Good
200
175
150
125
100
75
50
25
0
A73
A75
A74
A72
A71

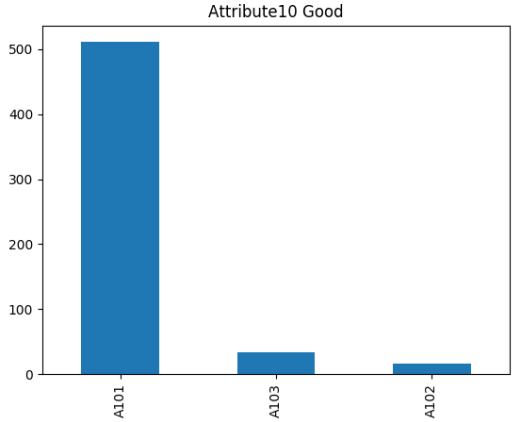

Attribute7 Bad
80
70
60
50
40
30
20
10
0
A73
A72
A75
A74
A71

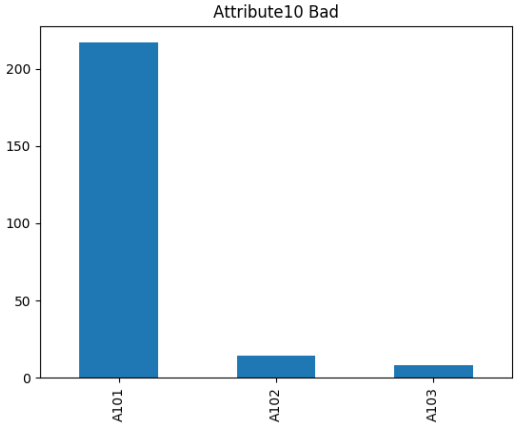

Attribute8 Good
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute8

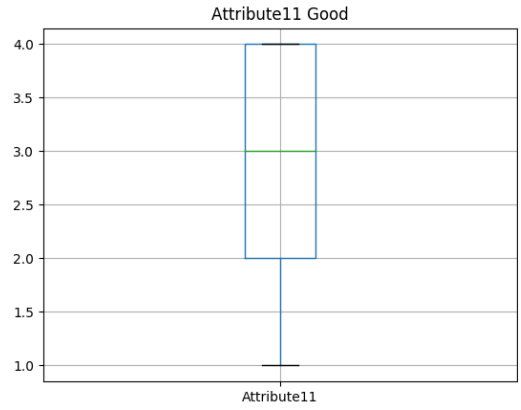

Attribute8 Bad
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute8

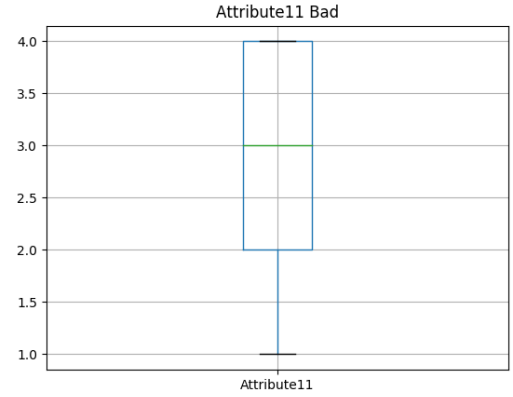

Attribute9 Good
300
250
200
150
100
50
0
A93
A92
A94
A91

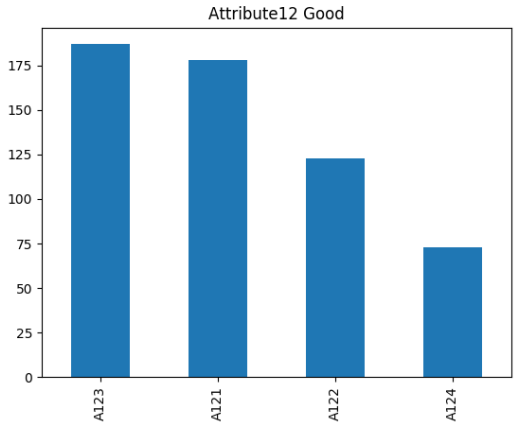

Attribute9 Bad
100
80
60
40
20
0
A93
A92
A94
A91

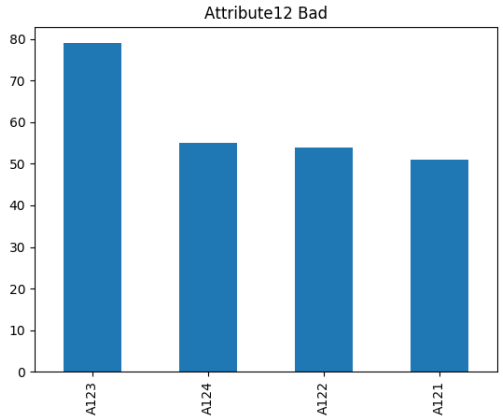

Attribute10 Good
500
400
300
200
100
0
A101
A103
A102

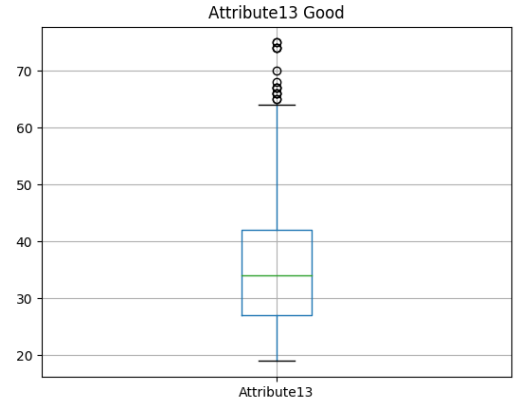

Attribute10 Bad
200
150
100
50
0
A101
A102
A103

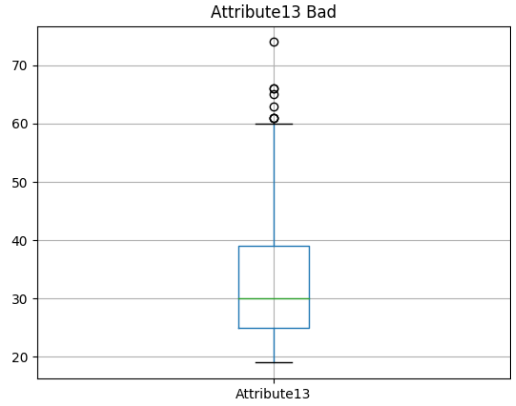

Attribute11 Good
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute11

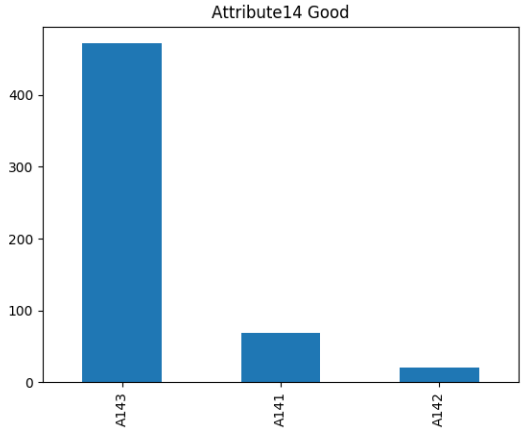

Attribute11 Bad
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute11

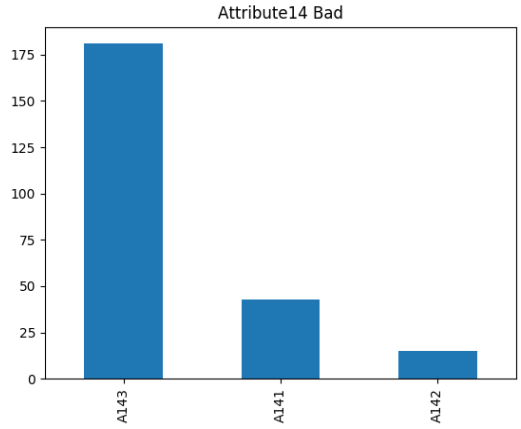

Attribute12 Good
175
150
125
100
75
50
25
0
A123
A121
A122
A124

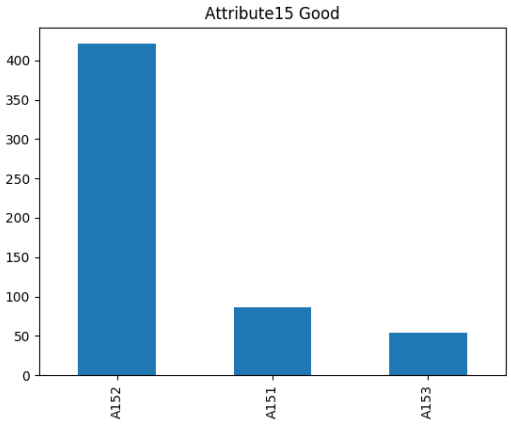

Attribute12 Bad
80
70
60
50
40
30
20
10
0
A123
A124
A122
A121

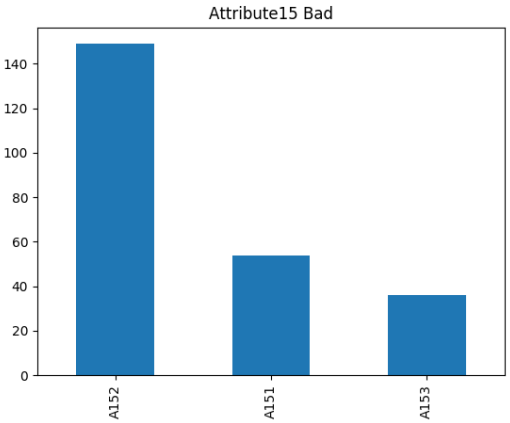

Attribute13 Good
70
60
50
40
30
20
Attribute13

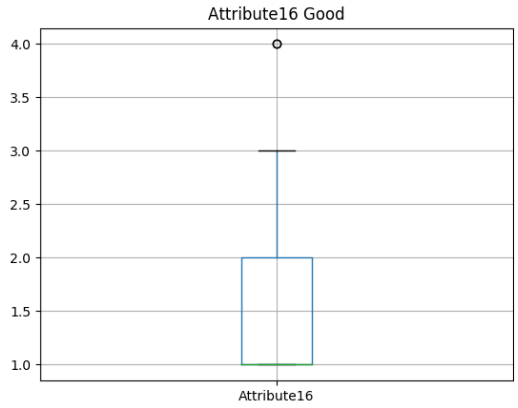

Attribute13 Bad
70
60
50
40
30
20
Attribute13

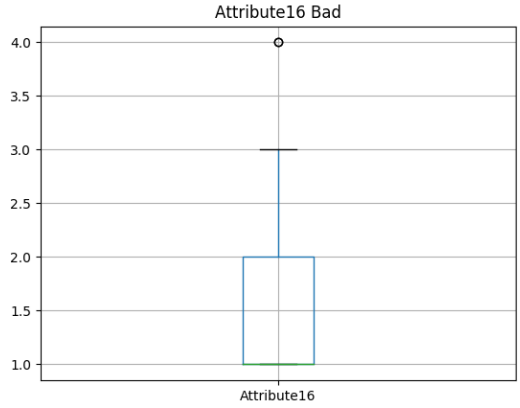

Attribute14 Good
400
300
200
100
0
A143
A141
A142

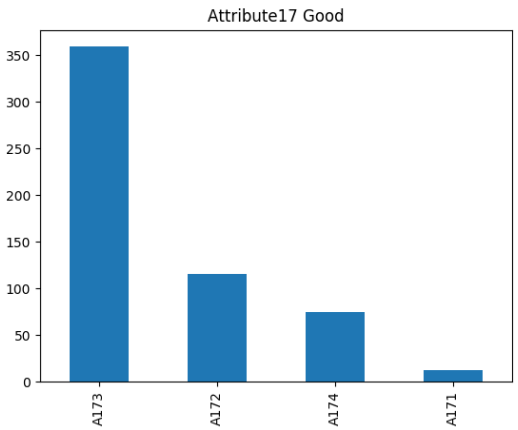

Attribute14 Bad
175
150
125
100
75
50
25
0
A143
A141
A142

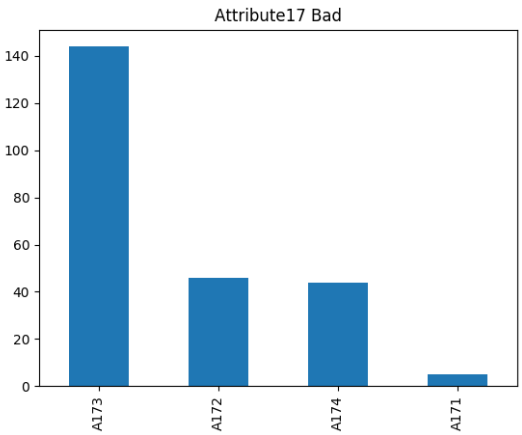

Attribute15 Good
400
350
300
250
200
150
100
50
0
A152
A151
A153

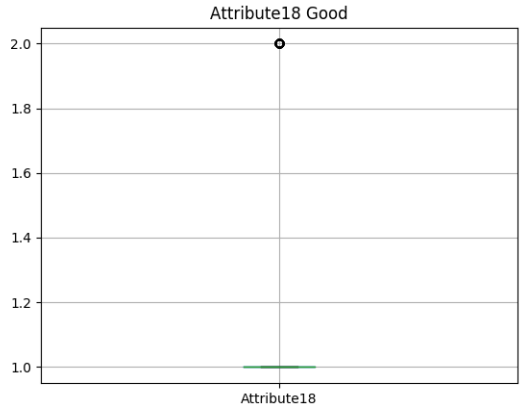

Attribute15 Bad
140
120
100
80
60
40
20
0
A152
A151
A153

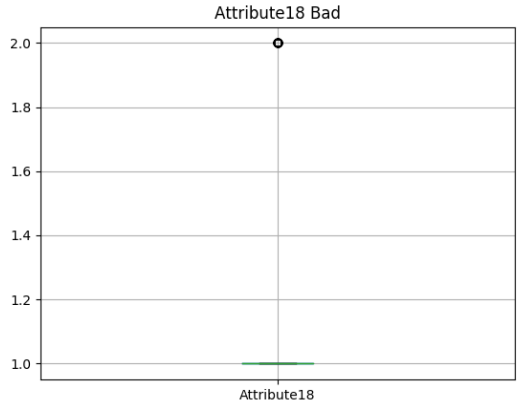

Attribute16 Good
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute16

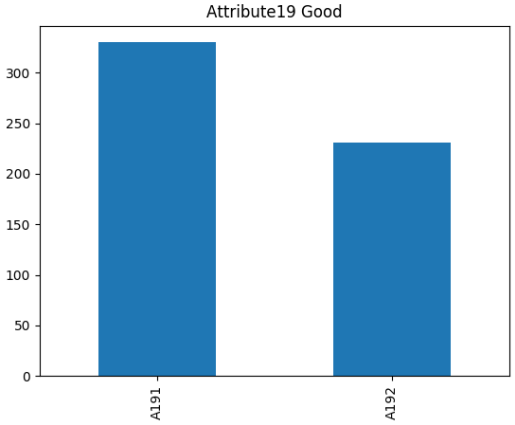

Attribute16 Bad
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Attribute16

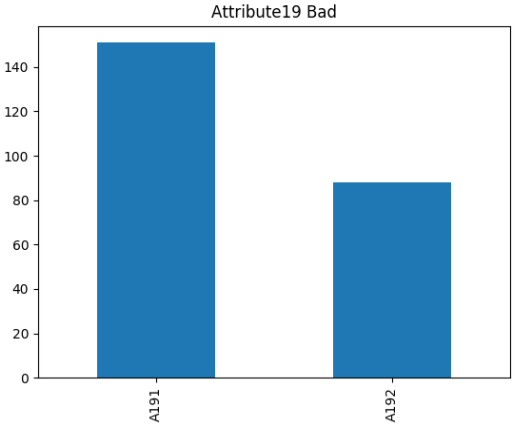

Attribute17 Good
350
300
250
200
150
100
50
0
A173
A172
A174
A171

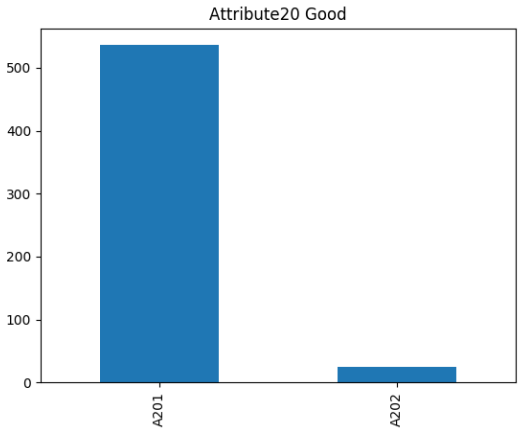

Attribute17 Bad
140
120
100
80
60
40
20
0
A173
A172
A174
A171

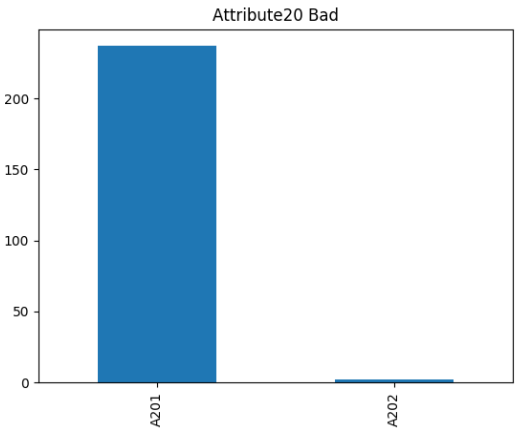

Attribute18 Good
2.0
1.8
1.6
1.4
1.2
1.0
Attribute18


Attribute18 Bad
2.0
1.8
1.6
1.4
1.2
1.0
Attribute18


Attribute19 Good
300
250
200
150
100
50
0
A191
A192


Attribute19 Bad
140
120
100
80
60
40
20
0
A191
A192


Attribute20 Good
500
400
300
200
100
0
A201
A202


Attribute20 Bad
200
150
100
50
0
A201
A202

Παρατηρούμε στα παραπάνω διαγράμματα ότι σε κάποια attributes τα γραφήματα έχουν ελάχιστες διαφορές μεταξύ good και bad ενώ σε κάποια άλλα έχουν σημαντικές διαφορές. Πιστεύουμε ότι όσο πιο μεγάλες οι διαφορές ανάμεσα στα διαγράμματα, τόσο πιο χρήσιμη η πληροφορία που προσφέρει το συγκεκριμένο attribute στην κατηγοριοποίηση των δανειοληπτών.

Πιο συγκεκριμένα έντονες διαφορές παρατηρούνται στα attributes: 1, 2, 3, 5, 9, 12, 13 και 15. Αντιθέτως τα attributes: 8, 10, 11, 16, 18 και 19 έχουν αμελητέες διαφορές. Αναμένουμε τα attributes με τις μεγαλύτερες διαφορές να έχουν και μεγαλύτερο information gain.

Classification

Εκτελέσαμε τους κατηγοριοποιητές Naive Bayes, Random Forest και SVM στο train.tsv με 10 fold cross validation και τα αποτελέσματα τα οποία προέκυψαν είναι τα εξής:

Statistic Measure	Naive Bayes	Random Forest	SVM
Accuracy	0.6375	0.74625	0.70125

Επιλογή Features

Με βάση τα προηγούμενα αποτελέσματα συμπεραίναμε ότι ο Random Forest είναι η καλύτερη επιλογή για την κατηγοριοποίηση των πελατών. Επομένως το αρχείο testSet_Predictions.csv συμπληρώθηκε με τη χρήση του παραπάνω κατηγοριοποιητή.

Στον παρακάτω πίνακα παρουσιάζεται το information gain σε αύξουσα ταξινόμηση για κάθε attribute.

Attribute	Information Gain
Attribute20	0.21
Attribute10	0.52
Attribute18	0.6
Attribute14	0.83
Attribute19	0.97
Attribute16	1.12
Attribute15	1.14
Attribute17	1.41
Attribute9	1.52
Attribute3	1.71
Attribute6	1.72
Attribute8	1.81
Attribute1	1.82
Attribute11	1.85
Attribute12	1.95
Attribute7	2.15
Attribute4	2.69
Attribute2	3.73
Attribute13	5.26
Attribute5	9.52

Όντως τα γραφήματα με μεγάλες διαφορές ανάμεσα σε good και bad στο προηγούμενο ερώτημα βρίσκονται χαμήλα στον πίνακα, δηλαδή έχουν μεγάλο Information Gain. Αντίθετα τα γραφήματα με μικρές διαφορές έχουν μικρό Information Gain.

Τέλος παρουσιάζεται η διακύμανση του μέσου accuracy για 10 fold cross validation στον κατηγοριοποιητή Random Forest καθώς αφαιρούνται feature. Αξίζει να σημειωθεί ότι η σειρά με την οποία αφαιρούνται τα features ακολουθούν την σειρά με την οποία είναι στον προηγούμενο πίνακα, δηλαδή αφαιρούνται πρώτα τα attributes με το μικρότερο information gain.